A Dimension Reduction Approach to Classification Based on Particle Swarm Optimisation and Rough Set Theory

Liam Cervante[1], Bing Xue[1], Lin Shang[2], and Mengjie Zhang[1]

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

Abstract. Dimension reduction aims to remove unnecessary attributes from datasets to overcome the problem of "the curse of dimensionality", which is an obstacle in classification. Based on the analysis of the limitations of the original rough set theory, we propose a new dimension reduction approach based on binary particle swarm optimisation (BPSO) and probabilistic rough set theory. The new approach includes two new specific algorithms, which are *PSOPRS* using only the probabilistic rough set in the fitness function and *PSOPRSN* adding the number of attributes in the fitness function. Decision trees, naive Bayes and nearest neighbour algorithms are employed to evaluate the classification accuracy of the reduct achieved by the proposed algorithms on five datasets. Experimental results show that the two new algorithms outperform the algorithm using BPSO with original rough set and two traditional dimension reduction algorithms. PSOPRSN obtains a smaller number of features than PSOPRS with the same or slightly worse classification performance. This work represents the first study on probabilistic rough set theory for for filter dimension reduction in classification problems.

Keywords: Dimension reduction, Particle Swarm Optimisation, Filter Approaches, Classification.

1 Introduction

Classification is an important task in machine learning and data mining. However, it often involves a large number of attributes in the datasets. The large attribute dimension causes the problem of "the curse of dimensionality" [1]. Dimension reduction, also called attribute reduction, aims to reduce the unnecessary attributes to reduce the attribute dimension while preserving the classification power of original attributes to maintain the classification performance [2]. By removing the unnecessary attributes, dimension reduction can reduce the training time of a learning algorithm and simplify the learnt classifier [3, 4].

Existing dimension reduction algorithms can be broadly classified into two categories: wrapper approaches and filter approaches [3, 5]. Wrapper approaches include a learning algorithm as part of the evaluation function to determine the goodness of the reduct. Therefore, wrappers can often achieve better results than filters [6]. Filter approaches are independent of a learning algorithm. Therefore, they are argued to be computationally cheaper and more general than wrappers.

Dimension reduction is a difficult task, where the size of the search space grows exponentially along with the number of attributes in the dataset. Although many different search techniques have been applied to dimension reduction, most of these algorithms suffer from the problems of stagnation in local optima or being computationally expensive [3, 7]. In order to better address dimension reduction problems, an efficient global search technique is needed. Evolutionary computation (EC) techniques are well-known for their global search ability. Particle swarm optimisation (PSO) [8, 9] is a relatively recent EC technique, which is computationally less expensive than other EC algorithms. Therefore, PSO has been used as an effective technique in dimension reduction [4, 10, 11].

EC algorithms (including PSO) have been successfully applied to address dimension reduction problems. However, most of the existing EC based dimension reduction algorithms are wrapper approaches. Although wrappers can achieve better classification performance, the use of wrappers is limited in real-world applications because of the high computational cost. The development of EC based filter dimension reduction approaches still remains an open issue. On the other hand, rough set theory has been applied to attribute reduction [12]. However, original rough set has limitations [13]. Probabilistic rough set can overcome such limitations and from a theoretical point of view, Yao and Zhao [13] have shown that probabilistic rough set can be a good measure in dimension reduction, but its performance has not been reported.

1.1 Goals

The overall goal of this paper is to develop a PSO based filter dimension reduction approach to classification to reduce the number of attributes and achieve similar classification performance to that of using all original attributes. To achieve this goal, we develop a new filter dimension reduction approach (with three new algorithms) based on PSO and probabilistic rough set theory. The proposed two dimension reduction algorithms will be examined and compared with a filter algorithm using original rough set theory and two traditional algorithms on five different benchmark datasets. Specifically, we will investigate

- whether using PSO and *original* rough set theory can reduce the number of attributes and maintain the classification performance,
- whether using PSO and *probabilistic* rough set theory can further reduce the number of attributes without decreasing the classification performance,
- whether considering *the number of attributes* in the fitness function can further reduce the number of attributes and maintain the classification performance.

2 Background

2.1 Particle Swarm Optimisation (PSO)

PSO is an evolutionary computation technique inspired by social behaviours of birds flocking and fish schooling [8, 9]. In PSO, each candidate solution is represented as a particle in the swarm and PSO starts with a number of randomly generated particles. All the particles move in the search space to find the optimal solutions. During the movement, each particle (i.e., particle i) has a position and velocity, which are represented by vectors $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$ and $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$, respectively, where D is the dimensionality of the search space. A particle can remember the best positions it visits so far, which is called personal best *pbest*. The best position obtained by the population thus far is called *gbest*, based on which a particle can share information with its neighbours. A particle iteratively updates its position and velocity to search for the optimal solutions based on *pbest* and *gbest* according to the following equations:

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{1}$$

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * (p_{id} - x_{id}^{t}) + c_2 * r_2 * (p_{gd} - x_{id}^{t}) \tag{2}$$

where t represents the tth iteration in the evolutionary process. $d \in D$ represents the dth dimension in the search space. w is the inertia weight, which can balance the local search and global search abilities of the algorithm. c_1 and c_2 are acceleration constants. r_1 and r_2 are random constants uniformly distributed in [0, 1]. p_{id} and p_{gd} denote the values of *pbest* and *gbest* in the dth dimension. v_{id}^{t+1} is limited by a predefined maximum velocity, v_{max} and $v_{id}^{t+1} \in [-v_{max}, v_{max}]$.

In order to extend PSO to address discrete problems. Kennedy and Eberhart [14] developed a binary particle swarm optimisation (BPSO). In BPSO, x_{id}, p_{id} and p_{gd} are restricted to 1 or 0. The velocity is still updated according to Equation (2), but it indicates the probability of the position in the corresponding dimension taking value 1. BPSO updates the position of each particle according to the following formula:

$$x_{id} = \begin{cases} 1, \text{ if } rand() <= \frac{1}{1+e^{-v_{id}}} \\ 0, \; otherwise \end{cases} \tag{3}$$

where $rand()$ is a random number selected from a uniform distribution in [0,1].

2.2 Rough Set Theory

Rough set theory developed by Pawlak [15] is a mathematical tool, which is able to deal with uncertainty, imprecision and vagueness. The main advantage of rough set theory is that it does not need any prior knowledge about the data.

In rough set theory, an information system can be denoted as $I = (U, A)$, where U is the universe of objects in the system and A is the set of attributes that describe each object. Equivalence relation is a relation that partitions a set so that every element of the set is a member of one and only one cell of the partition. Based on all equivalence relations described by A, the equivalence class relation partitions of U is $U_1, U_2, U_3, ..., U_n$, where n is the number of classes that objects in U may belong to.

For any $P \subseteq A$ and $X \subseteq U$, the equivalence relation is defined as $IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P, a(x) = a(y)\}$. An equivalence class of $IND(P)$ are denoted as $[x]_P$, which means that $\forall y \in [x]_P$ (x, y) are indiscernible with regards to

P. Based on the equivalent classes described by P, rough set theory defines a lower approximation ($\underline{P}X$) and an upper approximation ($\overline{P}X$) of X [15], where $\underline{P}X = \{x \in U | [x]_P \subseteq X\}$ and $\overline{P}X = \{x \in U | [x]_P \cap X \neq \emptyset\}$. $\underline{P}X$ contains all the objects, which can be surely classified to the target set X. $\overline{P}X$ contains the objects, which probably belong to the target set X.

An ordered pair $(\overline{P}X, \underline{P}X)$ is called a rough set. The concept of the reduct is fundamental for rough sets theory. A reduct is the essential part of an information system (related to a subset of attributes), which can achieve similar approximation power of classification as all the original attributes A. There can be many different reducts in a rough set and attribute reduction aims to search for the smallest reduct.

In the original rough set theory [15], $\underline{P}X$ and $\overline{P}X$ were defined as two extreme cases in terms of the relationships between an equivalence class defined by P and the target set X. $\underline{P}X$ requires that the equivalence class is a subset of X while $\overline{P}X$ requires the equivalence class must have a non-empty overlap with X. However, the degree of their overlap is not taken into account, which will limit its applications. Therefore, researchers investigate probabilistic rough set theory to relax the definitions of the lower and upper approximations [13].

In probabilistic rough set theory, $\mu_P[x]$ (See Equation 4) is defined as a way to measure the fitness of a given instance $x \in X$.

$$\mu_P[x] = \frac{|[x]_P \cap X|}{|[x]_P|} \tag{4}$$

The lower approximation is defined as Equation 5.

$$\underline{apr}_P X = \{x | \mu_P[x] \geq \alpha\} \tag{5}$$

where α can be adjusted to restrict or relax the lower and upper approximations. Note that $\underline{apr}_P X = \underline{P}X$ when $\alpha = 1$. $\underline{apr}_P X$ loosens the boundaries of the rough set. In a given equivalence class, if a large number of instances are in the target set X, but a small number of instances are not, $\underline{apr}_P X$ will include them in the lower approximation.

From theoretical point of view, Yao and Zhao have claimed that probabilistic rough set can be a good way for attribute reduction problems [13]. However, it has not been proved by any experiment.

2.3 Related Work on Dimension Reduction

A number of dimension reduction algorithms have been proposed in recent years [3, 1, 10]. Typical dimension reduction algorithms are reviewed in this section.

Traditional Dimension Reduction Approaches. A traditional filter dimension reduction approach is principal components analysis (PCA), which constructs a low-dimensional representation of the data by finding a few orthogonal linear combinations of the original variables with the largest variance [16]. Due to its conceptual simplicity and being relatively efficient, PCA has been widely used in practice. However, PCA increases the dimensionality of the data in some cases. Decision trees (DT) use only relevant attributes that are required to completely classify the training set and remove all other attributes. Cardie [17] proposes a filter based dimension reduction algorithm that uses a decision tree algorithm to remove unnecessary attributes for a nearest neighbourhood algorithm.

Two commonly used wrapper methods are SFS [18] and SBS [19]. SFS (SBS) starts with no attributes (all attributes), then candidate attributes are sequentially added to (removed from) the initial attribute subset until the further addition (removal) does not increase the classification performance. However, both SFS and SBS suffer from the problem of nesting effect, because an attribute is selected (eliminated) it cannot be eliminated (selected) later, which is so-called nesting effect [7]. The "plus-l-take away-r" method proposed by Stearns [20] could overcome this limitation by performing l times forward selection followed by r times backward elimination. However, the determination of the optimal values of (l, r) is a difficult problem.

EC Algorithms for Dimension Reduction. Evolutionary computation techniques have been applied to address dimension reduction problems, such as GAs, GP, ant colony optimisation (ACO) and PSO.

Based on GAs, Chakraborty [21] proposes a dimension reduction algorithm using a fuzzy sets based fitness function. However, PSO with the same fitness function in [22] achieve better performance than this GA based algorithm. Kourosh and Zhang [23] propose a dimension reduction algorithm using GP and naïve bayes (NB), where GP is used to combine attribute subsets and a set of operators together to find the optimal attribute subset. Ming [24] proposes a dimension reduction method based on ACO and rough set theory. Experimental results show that the proposed algorithm achieves better classification performance with fewer attributes than a C4.5 based dimension reduction algorithm.

As an EC technique, PSO has recently gained more attention for solving dimension reduction problems. Wang et al. [12] propose a filter dimension reduction algorithm based on an improved BPSO and rough set. However, the classification performance of the reduct was only tested on one learning algorithm, the LEM2 algorithm, which originally is specific used for rough set and has some bias for the proposed algorithm. Meanwhile, only using one learning algorithm can not show the advantage that filter algorithms is more general. Mohemmed et al. [11] propose a hybrid method (PSOAdaBoost) that incorporates PSO with an AdaBoost framework for face detection. PSOAdaBoost aims to search for the best attribute subset and determine the decision threshold of AdaBoost simultaneously, which speeds up the training process and increase the accuracy of weak classifiers in AdaBoost.

Chuang et al. [5] apply the so-called catfish effect to PSO for dimension reduction, which is to introduce new particles into the swarm by re-initialising the worst particles when *gbest* has not improved for a number of iterations. The introduced catfish particles could help PSO avoid premature convergence. Liu et al. [10] introduce a multi-swarm PSO (MSPSO) algorithm to search for the optimal attribute subset and optimise the parameters of SVM simultaneously. Experiments show that MSPSO could achieve higher classification accuracy than grid search, standard PSO and GA. However, MSPSO is computationally more expensive than the other three methods because of the large population size and complicated communication rules between different subswarms. Based on PSO, Unler and Murat [4] propose a dimension reduction algorithm with an adaptive selection strategy, where an attribute is chosen not only according to the likelihood calculated by PSO, but also to its contribution to the attributes already selected. Experiments suggest that the proposed method outperforms the tabu search and scatter search algorithms.

PSO has been shown to be an efficient search technique for dimension reduction by many existing studies. However, most of the existing approaches are wrappers, which are computationally more expensive and less general than filter approaches. Therefore, investigation of an effective PSO based filter dimension reduction algorithm is still an open issue. Probabilistic rough set was claimed to be a good way for dimension reduction problems [13], but its real performance has not been investigated. Therefore, it is thought to investigate the performance of probabilistic rough set and PSO for filter dimension reduction.

3 Proposed Filter Based Methods

Base on rough set theory and BPSO, we will propose a filter dimension reduction approach. Firstly, we use original rough set theory and BPSO for dimension reduction to see whether it can achieve good results. Then, we will develop a new approach based on probabilistic rough set theory and BPSO to further reduce the dimensionality.

3.1 BPSO and Original Rough Set Theory for Dimension Reduction(PSORS)

When using rough set theory for dimension reduction, the datasets for a classification problem can be regarded as an information system $I = (U, A)$, where all available attributes can be considered as A in the rough set theory. Based on the equivalence described by A, U can be partitioned to $U_1, U_2, U_3, ..., U_n$, where n is the number of classes in the dataset. After dimension reduction, the achieved reduct can be considered as $P \in A$. Therefore, the fitness of P can be evaluated by how well P represents each target set in U, which is a class in the dataset.

For $U_1 \in U$, let $\underline{P}U_1 = \{x \in U | [x]_P \subseteq U_1\}$ be the lower approximation of P according to U_1 if $[x]_P$ only contains instances in U_1. Let $\overline{P}U_1 = \{x \in U | [x]_P \cap U_1 \neq \emptyset\}$ be the upper approximation of P according to U_1 if $[x]_P$ contains at least one element not in U_1. The rough set, $\overline{P}U_1 - \underline{P}U_1$, contains every instance in U_1, but $\overline{P}U_1$ contains instances from other classes that are indiscernible with instances in U_1. Therefore, the purity of $[x]_P$ according to U_1 can be measured by $\frac{\underline{P}U_1}{\overline{P}U_1}$, which shows how well P represents the target set U_1. Therefore, how well P describe each target in U can be calculated by Equation 6, which is the fitness function in PSORS:

$$Fitness_1(P) = \frac{\sum_{U_i \in U} |PU_i|}{|\mathbb{U}|} \tag{6}$$

If the dimension reduction algorithm achieves a reduct with $Fitness_1(P) = 1.0$, it means the reduct can completely separate each class from other classes.

3.2 New Dimension Reduction Algorithm 1 (PSOPRS): Based on Probabilistic Rough Set Theory

As discussed in Section 2.2, the definitions of lower approximation and upper approximation limit the application of rough set theory. In classification problems, it may happen that two or more instances might have the same attribute values but be classified in different classes. This is possibly because incorrect values are entered or one instance is an exception to a class. Therefore, it is impossible to achieve the $Fitness_1(P) = 1.0$ in Equation 6. A set of attributes could be adequate, but erroneous or unusual values prevent these attributes being included in a reduct. This problem can be addressed by relaxing the definitions of lower and upper approximations in probabilistic rough set theory. Therefore, we propose a new filter attributes reduction algorithm (PSOPRS) based on BPSO and probabilistic rough set theory [25].

In PSOPRS, for the target set U_1, $\mu_P[x] = \frac{|[x]_P \cap U_1|}{|[x]_P|}$, which quantifies the proportion of $[x]_P$ is in U_1. Here $[x]_P$ does not have to be completely contained in U_1. $\underline{apr}_P U_1 = \{x | \mu_P[x] \geq \alpha\}$ defines the lower approximation of P according to U_1, where α can be adjusted to restrict or relax the lower or upper approximations. When $\alpha = 1.0$, $\underline{apr}_P U_1 = \underline{P} U_1$. The fitness function of PSOPRS is shown by Equation 7.

$$Fitness_2(P) = \frac{\sum_{x=1}^{n} |\underline{apr}_P X_i|}{|\mathbb{U}|} \tag{7}$$

3.3 New Dimension Reduction Algorithms 2 (PSOPRSN): Based on Probabilistic Rough Set Theory and Size of the Reduct

In PSOPRS, although the use of probabilistic rough can avoid the problems caused by original rough set, the number of attributes is not considered in the fitness function (Equation 7). For the same α value, if there are more than one reducts that have the same value of $Fitness_2(P)$, PSOPRS will not have the intention to search for the smaller reduct. Therefore, we propose a new algorithm, which searches for a reduct with the two objectives of maximising the representation power of the reduct (represented by $Fitness_2(P)$) and minimising the number of attributes in the reduct. A straightforward way to achieve this goal would be adding one component in fitness function $Fitness_2(P)$ to represent the number of attributes of the reduct, which is shown as Equation 8 and this method is called PSOPRSN:

$$Fitness_3(P) = \gamma * \frac{\sum_{x=1}^{n} |\underline{apr}_P X_i|}{|\mathbb{U}|} + (1-\gamma) * (1 - \frac{\#attributes}{\#totalAttributes}) \tag{8}$$

where $\gamma \in (0, 1]$ shows the relative importance of the representation power of the reduct while $(1 - \gamma)$ shows the relative importance of the number of attributes. As the range of $\frac{\sum_{x=1}^{n} |\underline{apr}_P X_i|}{|\mathbb{U}|}$ is in [0, 1], the number of attributes is converted to $(1 - \frac{\#attributes}{\#totalAttributes})$ to make sure the two components in the same ranges.

In PSORS and the two newly proposed algorithms, PSOPRS and PSOPRSN, the dimensionality of the search space is the number of attributes included in

Table 1. Datasets

Dataset	#Attributes	#Classes	#Instances
Lymphography (Lymph)	18	4	148
Spect	22	2	267
Dermatology	33	6	366
Soybean Large	35	19	307
Chess	36	2	3196

the dataset. Each particle is encoded in a binary string, where the "1" means the corresponding attribute is included in the reduct while "0" means the corresponding attribute is removed.

4 Experimental Design

Five datasets in Table 1 are used in the experiments, which were chosen from UCI machine learning repository [26]. They have different numbers of attributes, classes and instances, which are used as representative samples of the problems that the proposed algorithms will address. Note that all the five datasets are categorical data because rough set theory only works on discrete values.

In the experiments, the instances in each dataset are randomly divided into two sets: 70% as the training set and 30% as the test set. The proposed algorithms firstly run on the training set to obtain a reduct. The classification performance of the achieved reduct will be evaluated by a learning algorithm on the unseen test set. As filter algorithms, the learning algorithm only runs on the test set. Almost all learning algorithms can be used here. In order to test the claim that filter dimension reduction methods are general, 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 evaluate the classification performance of the achieved reduct on the test set.

In all algorithms, the fully connected topology is used in BPSO, $v_{max} = 6.0$, the population size is 30 and the maximum iteration is 100. $w = 0.7298$, $c_1 = c_2 = 1.49618$. These values are chosen based on the common settings in the literature [9]. Each algorithm has been conducted for 30 independent runs.

In PSOPRS, in order to test how the value of α influence the dimension reduction performance, four different α values are used in the experiments, which are 1.0, 0.9, 0.8, and 0.75. All the α values are larger than 0.5, because the lower approximation in probabilistic rough set should have the majority (at least have half) of the instances that belong to the target set. In PSOPRSN, α is set as 0.75 and five different γ values are used in the experiments, which are 1.0, 0.9, 0.8, 0.75, 0.5, to represent the different relative importance of the number of attributes in the fitness function. When $\alpha = 1$ in PSOPRS and $\gamma = 1$ in PSOPRSN, PSOPRS and PSOPRSN become the same as PSORS. Therefore, the results of PSOPRS $\alpha = 1$ and PSOPRSN with $\gamma = 1$ are not presented in the next section. In order to further examine the performance of the proposed algorithms, two conventional filter feature selection methods (CfsF and CfsB) in Weka [27] are used for comparison purposes in the experiments and the classification performance is calculated by DT.

Table 2. Results of PSORS and PSOPRS with DT as the learning algorithm

Dataset	Chess		Dermatology		Lymph	
Method	AveSize	Ave±Std(Best)	AveSize	Ave±Std(Best)	AveSize	Ave±Std(Best)
All	36	0.985	33	0.828	18	0.755
PSORS	30.70	0.983±0.003(0.987)	21.00	0.860±0.048(0.975)	11.73	0.724±0.068(0.796)
PSOPRS						
$\alpha = 0.9$	30.70	0.984±0.002(0.987)	21.00	0.860±0.048(0.975)	11.73	0.724±0.068(0.796)
$\alpha = 0.8$	29.97	0.983±0.003(0.985)	21.00	0.860±0.048(0.975)	11.77	0.723±0.068(0.796)
$\alpha = 0.75$	30.30	0.985±0.001(0.987)	21.00	0.860±0.048(0.975)	11.77	0.723±0.068(0.796)
Dataset	Soybean		Spect			
Method	AveSize	Ave±Std(Best)	AveSize	Ave±Std(Best)		
All	35	0.819	22	0.809		
PSORS	21.53	0.803±0.046(0.872)	21.00	0.860±0.048(0.975)		
PSOPRS						
$\alpha = 0.9$	21.60	0.805±0.044(0.872)	17.30	0.806±0.022(0.843)		
$\alpha = 0.8$	21.67	0.805±0.044(0.872)	17.50	0.800±0.020(0.820)		
$\alpha = 0.75$	21.63	0.804±0.043(0.872)	15.57	0.818±0.008(0.820)		

5 Experimental Results and Discussions

5.1 Experimental Results of PSORS

Tables 5 shows the experimental results of PSOPRS and PSOPRS on the five datasets and DT, NB and 5NN were used for classification. Due to page limit, only the results of using DT for classification are presented here. In Table 5, "All" means that all of the available attributes are used for classification. "AveSize" means the average number of attributes selected in the 30 independent runs. "Ave", "Std" and "Best" represent the mean, the standard deviation and the best classification accuracy achieved by DT across the 30 independent runs.

According to Table 5, it can be seen that in most cases, PSORS reduced around one third of the available attributes. After dimension reduction, the classification performance achieved by DT is still the similar to that of using all attributes. In almost all datasets, the best classification performance achieved by three learning algorithms using the reduct are the better than using all available attributes. The results suggestion that PSORS based on BPSO and original rough set theory can be successfully used to reduce the dimensionality and also improve the classification performance in many cases.

5.2 Experimental Results of PSOPRS

According to Table 5, it can be seen that in most cases, the number of remained attributes decreases when α in PSOPRS reduces. In terms of the classification performance, for DT, all the reducts can achieve similar classification performance to using all attributes. Although the mean classification accuracy is slightly worse than using all attributes in some cases, the best accuracy is better than using all attributes in all cases. Compared with PSORS, PSOPRS can further reduce the number of attributes and maintain the classification performance, especially when $\alpha = 0.75$. The results suggests that by using probabilistic rough set to evaluate the fitness of the attributes, the algorithm can further reducing the number of remained attributes without reduce its classification performance. A smaller α means more relax on the lower and upper approximations, which usually can slightly remove more unnecessary attributes to further reduce dimensionality of the datasets.

Table 3. PSOPRSN with $\alpha = 0.75$

Dataset	γ	AveSize	DT	NB	5NN
			Ave±Std(Best)	Ave±Std(Best)	Ave±Std(Best)
Chess	0.9	12.63	0.977±0.001(0.979)	0.927±0.009(0.945)	0.872±0.054(0.953)
	0.8	8.97	0.972±0.013(0.977)	0.929±0.010(0.953)	0.846±0.062(0.925)
	0.75	7.73	0.961±0.019(0.977)	0.932±0.009(0.953)	0.821±0.114(0.921)
	0.5	4.93	0.931±0.013(0.938)	0.931±0.013(0.941)	0.602±0.198(0.892)
Dermatology	0.9	8.17	0.757±0.068(0.918)	0.816±0.056(0.943)	0.787±0.058(0.877)
	0.8	8.07	0.775±0.078(0.967)	0.799±0.056(0.959)	0.784±0.060(0.918)
	0.75	7.73	0.743±0.085(0.926)	0.786±0.064(0.910)	0.766±0.073(0.893)
	0.5	6.43	0.752±0.093(0.951)	0.783±0.075(0.959)	0.725±0.083(0.943)
Lymph	0.9	5.03	0.667±0.033(0.673)	0.776±0.004(0.796)	0.753±0.011(0.755)
	0.8	5.00	0.661±0.046(0.673)	0.777±0.005(0.796)	0.752±0.010(0.755)
	0.75	5.00	0.673±0.000(0.673)	0.776±0.000(0.776)	0.755±0.000(0.755)
	0.5	4.00	0.714±0.000(0.714)	0.816±0.000(0.816)	0.796±0.000(0.796)
Soybean	0.9	9.70	0.714±0.031(0.767)	0.756±0.036(0.824)	0.675±0.037(0.749)
	0.8	9.00	0.705±0.038(0.780)	0.745±0.041(0.846)	0.665±0.039(0.749)
	0.75	8.77	0.713±0.043(0.775)	0.747±0.031(0.811)	0.668±0.033(0.749)
	0.5	7.47	0.713±0.039(0.802)	0.761±0.042(0.833)	0.670±0.033(0.727)
Spect	0.9	13.97	0.820±0.000(0.820)	0.767±0.010(0.775)	0.818±0.010(0.831)
	0.8	8.97	0.799±0.017(0.820)	0.783±0.024(0.820)	0.834±0.021(0.843)
	0.75	7.07	0.798±0.012(0.831)	0.797±0.029(0.843)	0.805±0.040(0.843)
	0.5	4.63	0.786±0.026(0.843)	0.796±0.025(0.843)	0.739±0.248(0.843)

Table 4. Results of CfsF and CfsB with DT as the learning algorithm

Dataset	Chess		Dermatology		Lymph		Soybean		Spect	
Method	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy
CfsF	5	0.781	17	0.873	8	0.733	12	0.805	4	0.70
CfsB	5	0.781	17	0.873	8	0.733	14	0.854	4	0.70

5.3 Experimental Results of PSOPRSN

According to Table 5.2, with a smaller γ can reduce can achieve a smaller the number of attributes. The reason is that a smaller γ means the number of attributes in PSOPRSN is more important than a relatively large γ. Compared with PSORS and PSOPRS, PSOPRSN can significantly reduce the number of attributes although the classification performance is slightly worse in some cases.

The results also show that when the number of attributes is reduced, the classification performance also decreases in most cases. The reason could be that $Fitness_3$ does not consider the number of equivalence classes in the dataset. In rough set, a small number of attributes (e.g. 12) can describe a large number (2^{12}) of equivalence classes. The problem here is that 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, $Fitness_3$ may achieve a small reduct, but it will loss generality and performs badly on unseen test data.

5.4 Comparisons With Two Traditional Algorithms

Experiments using CfsF and CfsB for dimension reduction have been conducted using Weka and DT was used for classification. The results are shown in Table 5.4. Comparing the experimental results of the four rough set theory based algorithm in Tables 5 and 5.2 with the two traditional algorithms, it can be seen that in almost all cases, although CfsF and CfsB can achieve a smaller size of attributes, the classification performance of CfsF and CfsB are smaller or much smaller than the rough set theory based algorithms, PSORS, PSOPRS and PSOPRSN. In terms of the computational time, both our proposed algorithms and two traditional algorithms used a relatively short time (less than 5 minutes in most cases).

6 Conclusions

This paper developed a new approach using *probabilistic* rough set theory and BPSO to remove irrelevant and redundant features and maintain the classification performance achieved by using all features. This new approach includes two new algorithms, which are BPSO and probabilistic rough set theory (*PSOPRS*) and BPSO with probabilistic rough set theory by adding the number of attributes in the fitness function (*PSOPRSN*). The performance of three new algorithms were examined and compared to BPSO with *original* rough set theory (*PSORS*) and two traditional methods, CfsF and CfsB, on five datasets. In order to test the generality of the proposed algorithms, the achieved reduct was evaluated by three different learning algorithms for classification on the unseen test sets. Experimental results show that in most cases, the three proposed algorithms can be successfully used for dimension reduction and outperform PSORS and the two traditional algorithms. PSOPRSN can significantly reduce the number of attributes in the reduct although the classification performance is slightly reduced in many cases. The reason might be that PSOPRSN does not consider the number of equivalence classes in the dataset.

This work represents the first study that successfully uses BPSO with probabilistic rough set for dimension reduction. In future, we will consider the number of equivalence classes in the dataset to further reduce the number of attributes without decreasing the classification performance and investigate its performance for dimension reduction and attribute selection problems on more datasets with a larger number of attributes. We also intend to investigate multi-objective PSO and rough set based filter algorithms to better explore the Pareto front of non-dominated solutions in dimension reduction and attribute selection to provide more informative solutions for users.

Acknowledgments. This work is supported in part by the National Science Foundation of China (NSFC No. 61170180,61035003), the Key Program of Natural Science Foundation of Jiangsu Province, China (Grant No. BK2011005) and the Marsden Fund of New Zealand (VUW0806) and the University Research Fund of Victoria University of Wellington (200457/3230).

References

1. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recognition **43**(1) (2010) 5–13
2. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3** (2003) 1157–1182
3. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis **1**(4) (1997) 131–156

4. Unler, A., Murat, A.: A discrete particle swarm optimization method for feature selection in binary classification problems. European Journal of Operational Research **206**(3) (2010) 528–539
5. Chuang, L.Y., Tsai, S.W., Yang, C.H.: Improved binary particle swarm optimization using catfish effect for feature selection. Expert Systems with Applications **38** (2011) 12699–12707
6. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97** (1997) 273–324
7. Yusta, S.C.: Different metaheuristic strategies to solve the feature selection problem. Pattern Recognition Letters **30** (2009) 525–534
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks. Volume 4. (1995) 1942–1948
9. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: IEEE International Conference on Evolutionary Computation (CEC'98). (1998) 69–73
10. Liu, Y., Wang, G., Chen, H., Dong, H.: An improved particle swarm optimization for feature selection. Journal of Bionic Engineering **8**(2) (2011) 191–200
11. Mohemmed, A., Zhang, M., Johnston, M.: Particle swarm optimization based adaboost for face detection. In: IEEE Congress on Evolutionary Computation (CEC'09). (2009) 2494–2501
12. Wang, X., Yang, J., Teng, X., Xia, W.: Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters **28**(4) (2007) 459–471
13. Yao, Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Information Sciences **178**(17) (2008) 3356 – 3373
14. Kennedy, J., Eberhart, R.: A discrete binary version of the particle swarm algorithm. In: IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation. Volume 5. (1997) 4104–4108
15. Pawlak, Z.: Rough sets. International Journal of Parallel Programming **11** (1982) 341–356
16. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics **2**(4) (2010) 433–459
17. Cardie, C.: Using decision trees to improve case-based learning. In: Proceedings of the Tenth International Conference on Machine Learning (ICML). (1993) 25–32
18. Whitney, A.: A direct method of nonparametric measurement selection. IEEE Transactions on Computers **C-20**(9) (1971) 1100–1103
19. Marill, T., Green, D.: On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory **9**(1) (1963) 11–17
20. Stearns, S.: On selecting features for pattern classifier. In: Proceedings of the 3rd International Conference on Pattern Recognition, Coronado, CA (1976) 71–75
21. Chakraborty, B.: Genetic algorithm with fuzzy fitness function for feature selection. In: ISIE'02. Volume 1. (2002) 315– 319
22. Chakraborty, B.: Feature subset selection by particle swarm optimization with fuzzy fitness function. In: ISKE'08. Volume 1. (2008) 1038–1042
23. Neshatian, K., Zhang, M.: Dimensionality reduction in face detection: A genetic programming approach. In: 24th International Conference Image and Vision Computing New Zealand (IVCNZ'09). (2009) 391–396
24. Ming, H.: A rough set based hybrid method to feature selection. In: International Symposium on Knowledge Acquisition and Modeling (KAM '08). (2008) 585–588
25. Yao, Y.: Probabilistic rough set approximations. Int. J. Approx. Reasoning **49**(2) (2008) 255–271
26. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann (2005)