International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 14053-14063

http://www.ripublication.com

Enhancing Semisupervised Clustering By Combining Rough Set Based Feature Reduction and Particle Swarm Optimization

Ms.V.R.Saraswathy[1], Dr. N.Kasthuri[2]
[1]Assistant Professor(SLG), Department of Electronics and Communication Engineering,
Kongu Engineering College, Tamilnadu, India. vrsaraswathy@kongu.ac.in
[2]Professor, Department of Electronics and Communication Engineering,
Kongu Engineering College, Tamilnadu, India. kasturi@kongu.ac.in

Abstract

In recent years, many applications often face the problem of curse of dimensionality. Increasing number of features in the clustering decreases the accuracy of clustering. Feature selection is necessary in many applications for effective information retrieval. The feature selection reduces the time consumption and memory wastage. The dataset may be imprecise, incomplete or uncertain. Rough sets deals with vagueness and uncertainty. Rough set theory (RST) has been successfully used as a selection tool to discover data dependencies and reduce the number of attributes contained in a dataset. Particle swarm optimization (PSO) is known to effectively solve large-scale nonlinear optimization problems. A semi-supervised hybrid feature selection based on PSO and RST for different datasets is proposed. Two feature selection algorithms namely PSO-quick reduct and PSO-relative reduct are applied for the different datasets. The simulation results of PSO-QR and PSO-RR show that hybridization of PSO with two rough set algorithms on semi-supervised data select features more effectively than rough set algorithms without hybridization of PSO.

Keywords: Feature selection, quick reduct, relative reduct, particle swarm optimization

Introduction

A. Feature Selection

Feature selection [12],[16],[15] approach aims to select a small subset of features that minimize redundancy and maximize relevance to the target (class labels). It can select

features relevant to a particular application, removes irrelevant and redundant features and improves the data quality.

Feature Selection techniques can be divided in two approaches: feature ranking and subset selection.

In the feature ranking approach, by ranking the features using some criteria and fixing threshold value, features above a defined threshold value are selected.

The second approach can be split in three parts:

1. ***Filter approaches:*** The features are selected first, then this subset is used to execute a classification algorithm.
2. ***Embedded approaches:*** The feature selection occurs as part a classification algorithm.
3. ***Wrapper approaches:*** An algorithm for classification is applied over the dataset in order to identify the best features.

B. Rough Set Theory

Rough set theory (RST) is one of the effective approaches used for the process of feature selection or feature reduction that uses the concept of approximations. RST [14],[13]has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information. A subset termed as reduct of the original attribute is obtained from the dataset with discretized attribute values using RST . All other attributes are removed from the dataset with minimal information loss**.**

The rough set itself is the approximation of a vague concept (set) with a pair of precise concepts, called lower and upper approximations [1]. These two are the classification of the data into different categories. The first (lower) approximation is a representation of the objects that are known with certainty to belong to the subset of interest whereas the second (upper) approximation is a representation of the objects which possibly belong to the subset.

Lower approximation set is known as positive region because X is characterized by a specific decision value. There is also a possibility of indiscernibility classes which contain only some tuples in X, that cannot be classified exactly. These are the objects in boundary region and mathematically, represented as $\underline{C}X$ - $C\square X$, where $C\square X$ is the upper approximation set. The elements belong to the negative region are obtained by subtracting U and $C\square X$, where X is a precise when $\underline{C}X$ is equal to $C\square X$, means boundary region is empty.

C. Particle Swarm Optimization

Particle Swarm Optimization is an evolutionary [4] ,[11] computation technique developed by Eberhart and Kennedy in 1995, which was inspired by the social behaviour of bird flocking and fish schooling. PSO [4] utilizes a “population” of particles that fly through the problem hyperspace with given velocities. After each iteration, the velocities of the individual particles are stochastically adjusted according to the historical best position for the particle itself and the neighbourhood best position.

By considering pbest, gbest and the velocity [6],[7],[8] of each particle, the update rule for their position is as the following equations:

$$V_{t+1}=w_t+C_1*rand()*(pbest-x_t)+C_2*rand()*(gbest-x_t) \quad (1)$$

$$x_{t+1}=x_t+V_{t+1} \quad (2)$$

where w is inertia weight which shows the effect of previous velocity vector (V_t) on the new vector V_{t+1}, C_1 and C_2 are acceleration constants and rand () is a random function in the range [0,1].x_t is current position of the particle and $x_{t+1}$is the new position of the particle.

Both the particle best and the neighbourhood best are derived according to a user defined fitness function. The each particle's motion naturally evolves to an optimal or near-optimal solution. PSO [6],[7],[8] is a computational intelligence-based technique that is not largely affected by the difficulty of size and nonlinearity problems, and easily provides the optimal solution in many problems where most analytical methods fail to converge.

D. Learning Methods

Supervised learning algorithms [8] are trained on labelled examples. The decision class is known for all the attributes.

Unsupervised learning algorithms [10],[14],[15] operate on unlabelled examples. The desired class is unknown. Here the goal is to find structure in the data, but not a mapping from inputs to outputs.

Semi-supervised learning [2] combines both labelled and unlabelled examples

Literature Review

In recent years, many applications have a large amount of data .The complexity involved in processing huge amount of data is very high. Hence feature selection is an important process to eliminate redundancy features and irrelevant features.

Sailesh Singh Panwar (2014) [12] proposed, "Feature Selection for High Dimensional Data". In this article, author has discussed about the Rough Set based feature selection algorithm called Quick Reduct which is an active research area in pattern recognition, data mining community and statistics. Rough Set tool greatly discover the data dependency and reduces attributes using data contained in dataset only. Hence, this feature selection technique has been used in high dimensional data to remove irrelevant features and this algorithm is not guaranteed to produce the minimal reduct. It fails for very high dimensional data.

C. Velayutham and K. Thangavel (2011)[15] proposed "Rough Set based Unsupervised feature selection using relative dependency measures". This work reports about Rough Set based Feature Selection algorithm namely Relative Reduct which is a backward elimination method. it is applied to both supervised and unsupervised data.

R.K Bania and B.Borah (2012)[1] proposed "A new modified Approach to Rough Set Feature Selection". This article reports about the fundamental Rough Set based

feature selection algorithms namely QR and RR. It also modified the QR algorithm using a stopping criterion with threshold and a concept of significance of features.

James Kennedy and Russell Eberhart (1995) [11] proposed "Particle Swarm Optimization". This work reports about the method for optimization of nonlinear functions. This optimization algorithm is better when compared to the genetic algorithm.

X.Wang, J.X.T. Yang, W. Xia, R. Jensen, (2007) [16] proposed," Feature selection based on rough sets and particle swarm optimization". The author used five different rough set algorithms. The Particle Swarms find optimal regions of the complex search space through the interaction of individuals in the population. PSO is suitable for selecting feature in that particle swarms will discover best feature combinations as they fly within the subset space.PSO does not need complex operators such as crossover and mutation, only primitive and simple mathematical operators are needed compared to GA. The experiment was conducted on UCI data and it was compared with GA-based approach and deterministic rough set reduction algorithms. The results prove that PSO is suitable for rough set-based feature selection.

H.Hannah Inbarani et al(2012)[6],[7] proposed, "Unsupervised Hybrid PSO – Quick Reduct Approach for Feature Reduction"and "Unsupervised Hybrid PSO – Relative Reduct Approach for Feature Reduction". These articles combine the benefits of both PSO and RST. PSO is an evolutionary computation technique which finds global optimum solution in many applications. These papers proposed a novel Unsupervised PSO based Quick Reduct (US-PSO-QR) and Unsupervised PSO based Relative Reduct (US-PSO-RR) for feature selection which employs a population of particles existing within a multi-dimensional space. The result is compared with the US-QR , US-RR algorithms respectively and time is reduced effectively.

H.Hannah Inbarani, Ahmad Taher Azarb, G.Jothi.(2013)[8] projected, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis", which focus on supervised feature selection methods based on hybridization of Particle Swarm Optimization (PSO), PSO based Relative Reduct (PSO-RR) and PSO based Quick Reduct (PSO-QR) are presented for the diseases diagnosis. The experimental result on several standard medical datasets proves the efficiency of this technique as well as enhancements over the existing feature selection techniques.

Daoqiang Zhang,Zhi-Hua Zhou and Songcan Chen (2007)[2] proposed, "Semi-Supervised Dimensionality Reduction". This paper deals with semi-supervised data which has both labeled and unlabeled decision classes. The simulation results prove that the proposed semisupervised approach is better than existing algorithms.

PSO Based Relative Reduct Algorithm

Relative reduct (RR) [13],[5],[9],[15] algorithm is a backward elimination method to reduce the irrelevant data in a given dataset. This is one of the most commonly used feature selection algorithm in Rough Set Theory (RST). It is easy and simple to implement. It selects features based on the relative dependency degree. The technique was originally proposed to avoid the calculation of discernability function [13] or positive regions, which can be computationally expensive without optimizations. The

main condition for RR [15] algorithm is the total dependency of the given dataset should be 1.

(A). The relative dependency measure [15] for the unsupervised data

$$\gamma_{R(a)} = \frac{|U/IND(R)|}{U/IND(R \cup \{a\})} \quad \forall\, a \notin R \qquad (3)$$

where a -attribute assumed to be eliminated from the dataset
R - Reduct subset of attributes.
IND - separation of records into similar and dissimilar subsets
U - total number of records

(B). The relative dependency measure for the supervised and semi-supervised data

$$\gamma_R(D) = \frac{|U/IND(R)|}{|U/IND(R \cup D)|} \qquad (4)$$

where D is the decision attribute.

(C). Pseudocode of PSO-RR

Input: C, the set of all conditional features;
D, the set of decision features;
Output: Reduct R
Step 1: Initialize X with random position and Vi with random velocity
$\forall$: $X_i \leftarrow$ random Position(); $V_i \leftarrow$ random Velocity(); fit $\leftarrow$ 0; globalbest $\leftarrow$ fit; Gbest $\leftarrow X_1$;
Pbest(1) $\leftarrow X_1$
For i=1. . .S
pbest(i)=X_i ;Fitness (i)=0;
End
For **Step 2** : While Fitness! = 1 //StoppingCriterion
For i = 1. . .S//for each particle
$\forall$: X_i Compute fitness of feature subset of X_i ;R={};
R$\leftarrow$Feature subset of X_i (1s of X_i) $\forall$ a $\in$ (C)
calculate γ_R
Fit = $\gamma_{R-\{a\}(D)}$; $\forall X \subset$ R;
End;
If Fitness (i) > fit ;Fitness (i) = fit ;Pbest(i) = X_i;
End If Fit == 1 return R;
End if ;
End For
For **Step 3**: Compute best fitness
For i=1:S

```
If (Fitness(i)>globalbest)
gloablbest ←Fitness(i); gbest←Xi;
End if ;
End For
Update Velocity(); //Update Velocity Vi of Xi
Update Position(); //Update position of Xi //Continue with the next iteration

End {while}
```

This algorithm calculates a reduct set without generating all the possible subsets. It selects the random values for each particle and velocity. A population of particles is generated with random positions and velocities on S dimensions in the problem space. For each particle in a population X_i, 1s are taken as the selected features and 0s are considered as features to be removed. The dependency of selected features is computed based on dependency of decision features. If the dependency is equal to1, then the feature subset related to the particle is considered as the reduct set. If the dependency is not equal to 1, pbest (highest relative dependency value) of each particle is retained and the best value of the entire population is retained as the global best value. Then the position and velocity are updated as defined above and the next population is generated and the fitness values are computed for each particle until fitness value of the selected feature subset becomes 1.

PSO Based Quick Reduct Algorithm

The Quick Reduct (QR)[1],[14] algorithm attempts to calculate a reduct without exhaustively generating all possible subsets. It starts with a null set R and adds subset of features after each iteration x. The subsets are added one at a time with features of greatest increase in the rough set dependency metric as given by equations (5) and (6). The process is repeated until the maximum value of 1is obtained for the dependency metric for the given dataset. In this algorithm, the dependency of each attribute is calculated and the best candidate is chosen.

(A)Dependency degree for supervised and semi-supervised data

The dependency degree can be calculated for the supervised and semi-supervised data by using the formula

$$\gamma_{P(Q)} = \frac{|POS_P(Q)|}{|U|} \tag{5}$$

where

POS - positive region.
U - Number of elements in a attribute or total number of records
P - Set of conditional attribute
Q - Decision attribute

(B). Dependency degree for unsupervised data

$$\gamma_{R\cup\{x\}}(y) = \frac{|POS_{R\cup\{x\}}(y)|}{|U|} \qquad (6)$$

where y -all the attributes of the dataset
R -reduct subset
U -total number of records
In unsupervised data[14], there is no decision class.

Supervised PSO based Quick Reduct Algorithm[8] computes a reduct set without generating all possible subsets. It starts with an empty set and it adds one at a time. A population of particles is constructed with random positions and velocities on S dimensions in the problem space. Each particle's position is represented as binary digits of length N, where N is the total number of attributes or features. Therefore, each particle's position is an attribute subset. Fitness function for each particle is evaluated.The process is same as that of PSO-RR, but fitness value is calculated by using (5) for supervised and semisupervised learning methods and using (6) for unsupervised learning method.

(C). Pseudocode of PSO-QR

Input: C, the set of all conditional features;
D, the set of decision features; Output: Reduct R
Step 1: Initialize X with random position and V_i withrandom velocity
$\forall$: $X_i \leftarrow$ random Position(); $V_i \leftarrow$ random Velocity();fit $\leftarrow$ 0; globalbest $\leftarrow$ fit;
Gbest $\leftarrow X_1$; Pbest(1) $\leftarrow X_1$,For i=1. . .S
pbest(i)=X_i ;Fitness (i)=0; End
For **Step 2** : While Fitness! = 1 //StoppingCriterion
For i = 1. . .S//for each particle $\forall$: X_i Compute fitnessof feature subset of X_i
R$\leftarrow$Feature subset of X (1s of X_i)
$\forall$ x $\in$ (C − R)
Fitness= $\frac{|POS_{R\cup\{x\}}(D)|}{|U|}$
Fit = Fitness ;$\forall X \subset$ R;End
For **Step 3**: Compute best fitness
For i=1:S If (Fitness(i)>globalbest) ;gloablbest$\leftarrow$Fitness(i);
gbest$\leftarrow X_i$; getReduct(X_i) Exit
End if ; End For
Update Velocity(); //Update Velocity V_i of X_i
Update Position(); //Update position of X_i //Continue with the next iteration
End {while}

Dataset Description

The five different datasets namely Erythemato Squamous data, Iris data, Ionosphere data, Liver disorder and Network data are used. The total number of features,

instances and decision classes for each dataset is given in Table 1. All the datasets are available in UCI repository database [3].

Table 1: Dataset Description

Dataset	**Instances**	**Features (Including decision attribute)**	**Decision classes**
Erythemato Squamous	366	34	6
Iris	150	5	3
Ionosphere	351	35	2
Liver Disorder	345	7	2

Results and Analysis

Based on RST, two feature selection algorithms namely QR and RR are hybridized with PSO .The feature selection algorithms are applied for the different datasets . The three learning methods namely supervised, unsupervised and semi-supervised are applied. Table 2 and 3 shows the comparison of different algorithms applied for Erythemato Squamous, Iris, Ionosphere and Liver Disorder taken from UCI repository.

(A) Quick Reduct Algorithm

Table 2: Result analysis of Quick Reduct Algorithm

	Algorithm	Erythemato Squamous		Iris Data		Ionosphere		Liver disorder	
		Features selected	Time (s)	Features selected	Time (s)	Features selected	Time(s)	Features selected	Time (s)
Without PSO	US-QR	19	4790	4	39.6	6	15349.5	3	389.6
	S-QR	13	125	4	3.47	3	817.27	3	19.12
	SS-QR	2	2.67	3	16.9	2	59.4	6	256
With PSO	US-PSO-QR	26	1941	4	68.3	15	5116.5	4	343.5
	S-PSO-QR	22	390	4	2.63	14	232.6	4	16.9
	SS-PSO-QR	17	370.7	4	3	19	211.5	4	15.2

(B) Relative Reduct Algorithm

Table 4: Result analysis of Relative Reduct Algorithm

	Algorithm	Erythemato Squamous		Iris Data		Ionosphere		Liver disorder	
		Features selected	Time (s)	Features selected	Time (s)	Features selected	Time (s)	Features selected	Time (s)
Without PSO	US-RR	15	50.8	4	5.37	6	37.7	3	10.34
	S-RR	10	41.88	3	6.9	3	60.68	3	10.07
	SS-RR	10	41.7	3	6.7	8	59.2	3	9.69
With PSO	US-PSO-RR	22	18.45	4	2.61	18	24.1	5	11.6
	S-PSO-RR	18	15.27	2	4.51	19	19.94	6	4.42
	SS-PSO-RR	18	14.89	2	4.59	13	18.17	4	4.5

RR and QR are applied directly to the datasets of different applications. More computational time is required for the unsupervised data which has no decision class when compared to algorithms hybridized with PSO. RR is easy and simple to evaluate compared to QR. RR selects features similar to QR but in a lesser time. Similarly, when QR without PSO is applied for semi-supervised data, it takes more time to analyze the features.

When the two feature selection algorithms are hybridized with PSO, its complexity is reduced in terms of processing time. In this case, compared to unsupervised and supervised data, these two algorithms are more suitable for the semi-supervised data. They have the benefits of both supervised and unsupervised learning methods. On hybridization with PSO also, RR is better than QR.

In table 2 and 3, the different algorithms based on QR and RR is applied for the Erythemato squamous, Iris, Ionosphere and Liver disorder datasets. For low dimensional datasets like Iris and Liver disorder, SS-PSO-RR and SS-PSO-QR take more time compared to S-PSO-RR and S-PSO-QR. Similarly for high dimensional datasets like Ionosphere and Erythemato squamous , SS-PSO based on QR and RR is better than others. From the table 2 and 3, it is clear that the proposed method is more suitable for high dimensional data. The simulation results show that in most of the cases SS-PSO-QR and SS-PSO-RR are better than other algorithms both in terms of feature selection and computational time.

Conclusion and Future Work

Recent advances in computing technology in terms of speed, cost, as well as the ability to process huge amounts of data in a reasonable time has motivated increased interest in feature selection to extract useful knowledge from data.

Feature selection is necessary to reduce redundancy to avoid the curse of dimensionality and to eliminate irrelevant features and reduce noise and to reduce time and space required. The simulation results show that the two RST algorithms namely QR and RR hybridized with PSO algorithm provides better results in terms of both time consumption and attribute reduction for Iris and Erythemato Squamous, Ionosphere and Liver disorder datasets when the decision attribute of the dataset is semi-supervised.

In future, RR and QR algorithms can be hybridized with other evolutionary algorithms and can be compared with PSO and the algorithms can be modified for dynamic datasets.

References

[1]. Bania R.K and Borah.B.,2012,"A new modified Approach to Rough Set Feature Selection", Proc. of the Intl. Conf. on Computer Applications – Volume 1.

[2]. Daoqiang Zhang, Zhi-Hua Zhou and Songcan Chen ,2007, "Semi-Supervised Dimensionality Reduction", Society of Industrial and Applied Mathematics.

[3]. Datasets: Available: https://archive.ics.uci.edu/ml/machine learning-databases.

[4]. Eberhart.R.C, Shi.Y.,2001,"Particle swarm optimization: developments, applications and resources", in: Proceedings of IEEE International Conference on Evolutionary Computation,Vol. 1,pp. 81-86.

[5]. Fu.X, Tan.F, Wang.H, Zhang.Y.Q, Harrison.R.,2006," Feature similarity based redundancy reduction for gene selection", Proceedings of the International Conference on Data Mining.

[6]. Hannah Inbarani .H, Nizar Banu P.K, Andrews S.,2012," Unsupervised Hybrid PSO – Quick Reduct Approach for Feature Reduction", Proceedings of IEEE Internatioal Conference on Recent Trends in Information Technology,Vol. 24,pp. 11-16.

[7]. Hannah Inbarani. H, Nizar Banu P.K, Andrews. S.,2012," Unsupervised Hybrid PSO – Relative Reduct Approach for Feature Reduction", Proceedings of IEEE Internatioal Conference on Recent Trends in Information Technology(ICRTIT),pp. 103-108.

[8]. Hannah Inbarani. H, Ahmad Taher Azarb. G, Jothi , 2013," Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis", ELSEVIER-Computer methods & programs in biomedicine, Vol. 3, pp. 175-185.

[9]. Hannah Inbarani. H, Nizar Banu.P.K.,2012," Performance Evaluation of Hybridized Rough Set based Unsupervised Approaches for Gene Selection", International Journal of Computational Intelligence and Informatics, Vol. 25,Issue 3-4,pp. 793-806.

[10]. Jothi. G, Hannah Inbarani.H.,2012 ,"Soft set based quick reduct approach for unsupervised feature selection", IEEE –International Conference on Advanced Communication Control and Computing Technologies.

[11]. Kennedy. J, Eberhart. R.C.,1995,"Particle swarm optimization", Proceedings of IEEE International Conference on Neural Networks.

[12]. Shailesh Singh Panwar ,Jan-2014,"Feature Selection for High Dimensional Data", International Journal of Enhanced Research in Science Technology & Engineering, Vol. 3 Issue 1, pp: (319-324).

[13]. Shampa Sengupta and Asit Kr. Das, 2012," Single Reduct Generation Based On Relative Indiscernibility Of Rough Set Theory", International Journal on Soft Computing (IJSC) Vol.3, No.1.

[14]. Velayutham. C, Thangavel.K.,2011," Unsupervised quick reduct algorithm using rough set theory", Journal of Electronic Science and Technology, Vol. 9,No.3pp. 193-201.

[15]. Velayutham.C,Thangavel.K.,2011,"Rough Set Based Unsupervised Feature Selection Using Relative dependency Measures", International Journal of Computational Intelligence and Informatics, Vol.1:No.1.

[16]. Wang.X,Yang.J.X.T, Xia.W, Jensen.R., 2007, "Feature selection based on rough sets and particle swarm optimization", Pattern Recognition Letters.