Particle Swarm Optimisation and Statistical Clustering for Feature Selection

Mitchell C. Lane, Bing Xue, Ivy Liu, and Mengjie Zhang

Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
{Mitchell C. Lane, Bing.Xue, Mengjie.Zhang}@ecs.vuw.ac.nz, Ivy.Liu@msor.vuw.ac.nz

Abstract. Feature selection is an important issue in classification, but it is a difficult task due to the large search space and feature interaction. Statistical clustering methods, which consider feature interaction, group features into different feature clusters. This paper investigates the use of statistical clustering information in particle swarm optimisation (PSO) for feature selection. Two PSO based feature selection algorithms are proposed to select a feature subset based on the statistical clustering information. The new algorithms are examined and compared with a greedy forward feature selection algorithm on seven benchmark datasets. The results show that the two algorithms can select a much smaller number of features and achieve similar or better classification performance than using all features. One of the new algorithms that introduces more stochasticity achieves the best results and outperforms all other methods, especially on the datasets with a relatively large number of features.

Keywords: Feature selection, Particle swarm optimisation, Statistical clustering

1 Introduction

A machine learning technique (e.g. a classification algorithm) often suffers from the problem of high dimensionality. Feature selection aims to select a small subset of relevant features to reduce the dimensionality, maintain or increase the classification performance and simplify the learned classifiers [1].

Feature selection is a difficult task due mainly to the large search space and the feature interaction problem [2]. Most of the existing methods suffer from the problem of stagnation in local optima. Particle swarm optimisation (PSO) [3, 4] is an arguable global search technique, which has been successfully applied to many areas, including feature selection [5, 6]. In PSO, a candidate solution is represented by a particle in the swarm. Particles fly in the search space to find the optimal solutions by updating the velocity and position of each particle. In binary PSO (BPSO) [7], each particle is encoded as a binary string (i.e. "1" and "0"). The velocity value represents the probability of the corresponding dimension in the position taking value "1". The detailed description of BPSO is not presented here due to the page limit and it can be seen in [7].

Many statistical measures have been applied to form the evaluation function in a feature selection algorithm [1, 8]. However, all of them are used in filter approaches, which can not achieve as good classification performance as wrapper approaches [1]. This paper uses a new statistical clustering method [9, 10] that groups relatively homogeneous features together based on a statistical model. The method considers all features simultaneously and takes the feature interaction into account. Features in the same cluster are similar and they are dissimilar to features in other clusters. Since the feature interaction is an important issue in feature selection, the statistical feature interaction information found by the clustering method can be used to develop a good feature selection algorithm. However, this has seldom been investigated.

1.1 Goals

The overall goal of this paper is to investigate the use of statistical clustering information in PSO for feature selection. To achieve this goal, a statistical clustering method as a preprocessing step is performed on part of the training set to group features to different clusters. A simple greedy forward search (GFFS) is developed to select one feature from each cluster and then two new PSO based algorithms are proposed to search for a better combination of features from each cluster. Specifically, we will investigate:

- whether the simple GFFS can effectively use the clustering information to select a small number of features and achieve similar or even better classification accuracy than using all features,
- whether PSO with the clustering information produced by statistical clustering can achieve better performance than GFFS, and
- whether the introduction of a greater amount of stochasticity to the above new PSO based algorithm can further improve the classification accuracy.

2 Proposed Feature Selection Approaches

In this work, we use a newly developed clustering method proposed by Pledger and Arnold [9] and Matechou et. al. [10], which is not described here due to the page limit. The clustering method is performed as a preprocessing step on a small number of training instances to cluster features into different groups. Features in the same cluster are considered as similar features. Selecting multiple features from the same cluster may bring redundancy. Features from different clusters are more likely to be complementary to each other, which can increase the classification performance. Therefore, we first develop a simple greedy forward selection algorithm to select a single feature from each cluster to investigate whether the selected features can obtain similar or better classification performance than using all features. We then propose two BPSO based feature selection algorithms to search for a better feature subset.

2.1 Greedy Forward Feature Selection (GFFS)

GFFS is proposed based on the idea of sequential forward selection, where features are sequentially added to the feature subset, but the key part of GFFS is the use of the statistical clustering information.

GFFS starts with an empty feature set S and features are sequentially added into S according to the classification performance. Each individual feature is first used for classification on the training set. Features are then ranked according to the classification accuracy. The highest ranked feature that has the best classification performance is added to the feature subset S and other features in the same cluster are removed. For the remaining features, the feature combined with which S can achieve better classification performance than with others is added to S. The other features in the same cluster are

Algorithm 1: Pseudo-code of PSOMP

```
begin
    initialise position $x$ and velocity $v$ of each particle,
    random select one feature from each cluster;
    while $Maximum\ Iterations$ has been not met do
        evaluate the classification performance of the selected features;
        update $pbest$ and $gbest$ of each particle;
        for i=1 to $Swarm\ Size$ do
            for d=1 to $Dimensionality$ do
                update $v_i$ ;                                        /* Update velocity */
            for C=1 to $Clusters\ Size$ do
                find the dimension ($LD$) with the largest velocity in the cluster $C$ ;   /* feature with
                the highest probability */
                update the position value in dimension $LD$ to 1 ;   /* Update position */
                update other dimensions(features) in $C$ to 0 ;      /* Update position */
    calculate the training and testing classification performance of the selected features;
    return $gbest$, the training and testing classification performance.
```

removed. This procedure is repeated until all clusters have been visited and only one feature is selected from each cluster. The number of features selected by GFFS is the number of feature clusters, which is much smaller than the total number of features.

2.2 BPSO for Feature Selection Based on Maximum Probability (PSOMP)

Traditionally, when using PSO for feature selection, features are selected from the whole feature set [5, 6]. In order to select one feature from each cluster, we first develop a new BPSO based feature selection algorithm named (PSOMP), where features are selected from each cluster according to the maximum probability calculated by BPSO.

When using PSO for feature selection, each feature corresponds to one dimension in the position and velocity. "1" in the position means the corresponding feature is selected and "0" otherwise. BPSO may select more than one feature from each cluster. Therefore, PSOMP is proposed to select a single feature from each cluster. When using clustering information, a cluster of features correspond to a number of dimensions. Selecting one feature from each cluster means only one of these dimensions in the position can be updated to "1". To achieve this, the maximum probability mechanism is developed in PSOMP, where the motivation is that the velocity in BPSO represents the probability of the corresponding dimension taking value "1" [7]. In terms of feature selection, the velocity represents the probability of a feature being selected, i.e. the feature with the highest velocity has the maximum probability to be selected. Therefore, PSOMP updates the position value of only one feature (with the highest velocity) to "1", and updates all the other position values in the same cluster to "0".

Algorithm 1 shows the pseudo-code of PSOMP. The classification performance of the selected features is used to form the fitness function in PSOMP. The number of features selected equals to the number of feature clusters.

2.3 PSOMP with Tournament Feature Selection (PSOTFS)

PSOMP is based solely upon the probability of each feature, which allows PSOMP to select a single feature from each cluster, but may result in the quick (premature) convergence of the swarm. To resolve this problem, a tournament feature selection operator is introduced to PSOMP to develop a new algorithm named PSOTFS.

Table 1. Datasets

Dataset	# Features	# Instances	# Classes	# **Clusters**
Australian Credit Approval (Aus.)	14	690	2	7
Vehicle	18	846	4	5
German	24	1000	2	10
World Breast Cancer Diagnostic (WBCD)	30	569	2	8
Lung Cancer	56	32	3	7
Sonar	60	208	2	10
Musk Version 1 (Musk1)	166	476	2	12

The goal of using the tournament selection operator is to introduce some stochasticity to the swarm to ensure the diversity of the population. Note that the tournament selection operator is not applied to the individual particles in PSO, but to the features in the same cluster to select a sub-group of features. The tournament selection operator is applied before the position updating procedure in Algorithm 1 (after Line 10). It randomly selects a sub-group of features from a feature cluster. Then the maximum probability mechanism (in Line 11) is applied on the selected sub-group (instead of on the whole cluster in PSOMP) to find the feature with the highest probability in the sub-group. The position value of this feature is updated to “1” and that of all features in the same cluster are updated to “0”. Algorithm 1 can show the pseudo-code of PSOTFS by adding the tournament selection after Line 10 and replacing the cluster in Line 11 with the sub-group selected by the tournament selection.

3 Experimental Design

Seven benchmark datasets (Table 1) were chosen from the UCI machine learning repository [11] to test the performance of the proposed algorithms, GFFS, PSOMP and PSOTFS. The number of clusters obtained from the statistical clustering method is listed in the last column of Table 1. The instances in each dataset are split randomly into a training set (70%) and a test set (30%). K-Nearest Neighbour (KNN) with K=5 is used to evaluate the classification performance of the selected features.

The parameters of in PSOMP and PSOTFS are set as follows: w = 0.7298, c_1 = c_2 = 1.49618, population size is 30, the maximum number of iterations is 100 and the fully connected topology is used. These values are chosen based on the common settings in the literature [4]. The size of the tournament feature selection in PSOTFS is half of the number of features in the cluster. On each dataset, GFFS obtained a unique solution because it is a deterministic algorithm. PSOMP and PSOTFS have been conducted for 50 independent runs on each dataset. Student’s T-tests (Z-tests) are performed to compare their classification performances, where the significance level was selected as 0.05.

4 Results and Discussions

Experimental results are shown in Table 2, where “T1” represents the results of the T-Test between the classification performance of each new algorithm and that of using all features. “T2” represents the results of the T-Test between the classification accuracy achieved by GFFS and that of PSOMP or PSOTFS.

Note that since each algorithm is only allowed to select a single feature from each cluster, the number of features selected by all the three algorithms is the same as the number of feature clusters. Therefore, each algorithm selected a significantly smaller number of features than the total number of features in the dataset.

Table 2. Experimental Results

Dataset	Method	NO. of Features	Accuracy				Dataset	Method	NO. of Features	Accuracy			
			Ave (Best)	Std	T1	T2				Ave (Best)	Std	T1	T2
Aus.	All	14	70.05				Vehicle	All	18	83.86			
	GFFS	7	70.53		+			GFFS	5	84.84		+	
	PSOMP	7	73.43 (73.43)	0	+	+		PSOMP	5	84.41 (84.84)	8.1E-3	+	-
	PSOTFS	7	73.43 (73.43)	0	+	+		PSOTFS	5	84.84 (84.84)	1E-015	+	=
German	All	24	68.33				WBCD	All	30	92.98			
	GFFS	10	68.67		+			GFFS	8	89.47		-	
	PSOMP	10	69.67 (72.00)	0.0047	+	+		PSOMP	8	93.91 (94.74)	0.00574	+	+
	PSOTFS	10	69.67 (69.67)	1E-015	+	+		PSOTFS	8	92.98 (92.98)	0	=	+
Lung	All	56	70				Sonar	All	60	76.19			
	GFFS	7	90		+			GFFS	10	76.19		=	
	PSOMP	7	80.2 (90.00)	0.0424	+	-		PSOMP	10	75.65 (82.54)	0.0322	=	=
	PSOTFS	7	80.8 (90.00)	0.0337	+	-		PSOTFS	10	76.29 (85.71)	0.0337	=	=
Musk1	All	166	83.92										
	GFFS	12	79.02		-								
	PSOMP	12	80.8 (86.01)	0.0233	-	+							
	PSOTFS	12	81.62 (87.41)	0.0266	-	+							

Results of GFFS. According to Table 2, on five of the seven datasets, GFFS maintained or improved the classification performance by using only the selected small number of features. On the WBCD and Musk1 datasets, although the classification performance of GFFS is slightly decreased, the number of features is significantly reduced. The results suggest that this simple greedy forward selection algorithm can utilise the information provided by the clustering method to effectively reduce the number of features and achieve similar or higher classification performance than using all features.

Results of PSOMP. Table 2 shows that PSOMP achieved significantly higher classification accuracy than using all features on five of the seven datasets and similar performance on one dataset. Although on Musk1, the average classification performance of PSOMP is slightly (3%) lower than using all features, PSOMP removed around 92% of the original features, which considerably reduced the classification time and dimensionality. Meanwhile, the best classification performance of PSOMP is 2% higher than using all features. The results suggest that PSOMP using the statistical information to guide the search of BPSO can successfully address feature selection problems.

PSOMP discovered feature subsets with significantly better or similar classification performance to GFFS in most cases. The results suggest that PSOMP using BPSO as the search technique can better search the solution space to obtain better results than GFFS. Meanwhile, rather than obtaining a single solution by GFFS, PSOMP can generate multiple results, which has a higher probability to achieve better performance.

Results of PSOTFS. According to Table 2, PSOTFS selected a significantly smaller number of features and achieved similar or significantly higher accuracy than using all features on six of the seven datasets. Only on the Musk1 dataset, the average classification accuracy of PSOTFS is around 2% lower than using all features, but its best accuracy is around 4% higher and it selected only around 7% of the original features. PSOTFS achieved similar or higher accuracy than GFFS on six of the seven datasets. Compared with PSOMP, PSOTFS achieved similar performance to PSOMP on five of the seven datasets, where the number of features is relatively small. On the Sonar and Musk1 datasets with a slightly larger number of features, PSOTFS outperformed PSOMP in terms of both the average and the best classification performance.

The results suggest that a greater amount of stochasticity in PSOTFS maintains the swarm diversity to avoid premature convergence. Therefore, PSOTFS achieved higher classification accuracy than PSOMP in most cases, especially on the datasets with a larger number of features and the solution space is more complex.

5 Conclusions

The goal of this paper was to investigate the use of statistical clustering methods in PSO for feature selection. The goal was successfully achieved by developing two new PSO based feature selection approaches, PSOMP and PSOTFS, to select a single feature from each cluster. The proposed algorithms are compared with a simple greedy forward feature selection algorithm (GFFS) on seven datasets. The experiments show that by using the statistical clustering information, GFFS selected a small number of features and achieved better classification performance than using all features. The basic PSOMP outperformed GFFS in most cases and PSOTFS achieved better classification performance than PSOMP because of the introduction of stochasticity to the swarm.

This study is a preliminary work of successfully using statistical clustering in feature selection, which motivates us to further investigate this research topic, such as using PSO to select multiple or zero features from each cluster to further improve the performance and using statistical clustering information for feature construction.

References

1. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis **1** (1997) 131–156
2. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3** (2003) 1157–1182
3. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks. Volume 4. (1995) 1942–1948
4. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: IEEE International Conference on Evolutionary Computation (CEC'98). (1998) 69–73
5. Xue, B., Zhang, M., Browne, W.: Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (2012) DOI:10.1109/TSMCB.2012.2227469
6. Wang, X., Yang, J., Teng, X., Xia, W.: Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters **28** (2007) 459–471
7. 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
8. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis. Technical report (2005)
9. Pledger, S., Arnold, R.: Multivariate methods using mixtures: correspondence analysis, scaling and pattern detection. Computational Statistics and Data Analysis (online: http://dx.doi.org/10.1016/j.csda.2013.05.013) (2013)
10. Matechou, E., Liu, I., Pledger, S., Arnold, R.: Biclustering models for ordinal data. Presentation at the NZ Statistical Assn. Annual Conference, University of Auckland (2011)
11. Bache, K., Lichman, M.: UCI Machine Learning Repository (2013)