Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 1, 2021, 39-44.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Feature Selection based on Particle Swarm Optimization and Mutual Information

Zahra Shojaee[1], Seyed Abolfazl Shahzadeh Fazeli[2*], Elham Abbasi[3] and Fazlollah Adibnia[4]

1,2,3. Department of Computer Science, Yazd University, Yazd, Iran.
4. Department of Computer Engineering, Yazd University, Yazd, Iran.

Article Info

Article History:
Received 08 December 2019
Revised 05 February 2020
Accepted 06 July 2020

DOI:10.22044/jadm.2020.8857.2020

Keywords:
Feature Selection, Mutual Information, PSO, Classification, Machine Learning.

**Corresponding author: fazeli@yazd.ac.ir (S.A. Shahzadeh Fazeli).*

Abstract

Today, feature selection, as a technique to improve the performance of the classification methods, has been widely considered by the computer scientists. As the dimensions of a matrix has a huge impact on the performance of processing on it, reducing the number of features by choosing the best subset of all the features. It will affect the performance of the algorithms. Finding the best subset by comparing all the possible subsets, even when n is small, is an intractable process, and hence, many research works have approached to the heuristic methods to find a near-optimal solutions. In this paper, we introduce a novel feature selection technique that selects the most informative features and omits the redundant or irrelevant ones. Our method is embedded in PSO (Particle Swarm Optimization). In order to omit the redundant or irrelevant features, it is necessary to figure out the relationship between different features. There are many correlation functions that can reveal this relationship. In our proposed method, to find this relationship, we use the mutual information technique. We evaluate the performance of our method on three classification benchmarks: Glass, Vowel, and Wine. Comparing the results obtained with four state-of-the-art methods demonstrates its superiority over them.

1. Introduction

In the machine learning techniques, classification problems have been considered by many researchers. In these problems, the objects are categorized into different classes according to their similarities or differences. The base of comparison between objects is their features. Each object is considered as a vector of characteristics, and will be compared with the other objects to be categorized in different classes.

Today, by the progress of information retrieve techniques and tools, datasets with a large number of features and relatively few patterns are produced. A large number of irrelevant or redundant features may significantly decrease the accuracy of the learned models as well as increasing the computational complexity of building the model. This problem is called curse of dimensions. As a solution, the feature selection techniques are designed to reduce the dimensionality of the datasets by selecting the most informative features without losing important information for the classification task. They omit irrelevant and redundant features. The irrelevant features can mislead us and the redundant features add no new knowledge. Feature selection can improve the classification accuracy, and also reduces the number of features.

Feature selection has many practical applications in different fields such as text categorization [1], face recognition [2], gene classification [3], cancer prediction [4], fraud detection [5], and recommender systems [6] . In order to find the optimal subset, one has to explore the power set of features whose running Time is $O(2^n)$, and hence, this is an intractable problem.

Hence, finding the optimal feature subset is computationally expensive and also impractical for even a moderate-sized feature set. Many feature selection algorithms involving heuristic techniques are presented to find the optimal or near optimal subset of features, the same as GA (Genetic Algorithm) [7] and GP (Genetic Programming) [8], PSO [9], ACO (Ant Colony Optimization) [10], memetic algorithms [11], and ABC (Artificial Bee Colony) [12]. ACO uses a graph to represent the search space such that features are encoded as nodes to construct a graph model. Each ant represents a feature subset [10]. In most ACO-based algorithms, nodes of the graph are fully connected. However, in [13], each feature was connected only to two other features. A binary set with length of the number of nodes that an ant will visit will be the final solution. In feature selection, the representation of each particle in PSO is a string, in which the length of string is equal to the number of features in the dataset. In the binary version, 1 and 0 represent the selecting and deselecting of the corresponding feature, respectively. In the continuous representation, where elements are the real-value numbers, a threshold α is usually used to determine the selection of a particular feature. If the value is larger than α, the corresponding feature is selected; otherwise, it is not selected. The length of the new representation is equal to the total number of features and parameters. The representation is encoded in three different ways: continuous encoding [14], binary encoding [15], and a mixture of binary and continuous encoding [16]. Also PSO has been applied to multi-objective filter feature selection, where information-based theory [17] and rough set theory [18] have been used to evaluate the relevance of the selected features. These works showed that PSO for multi-objective feature selection provided multiple solutions to the users. Our proposed method is based on PSO with the objective of omitting the irrelevant or redundant features according to their relationship.

In order to find the relationship between the features, we use the mutual information technique. We also implement our classification phase by applying the K-means algorithm.

To omit the irrelevant and redundant features, we need a criterion to evaluate the relationship of a candidate feature along with the already selected features. Finding a relationship between two random variables is called correlation in statistics. The correlation methods like Pearson and Spearman estimate the linear relationship. In other words, it cannot determine all the relevant features and it does not satisfy all of our necessities. Mutual information [19] provides a more powerful tool for determining the relationship of variables. It measures the reduction of uncertainty in X_2 after observing X_1. It can measure non-monotonic relationships and other more complicated relationships. Many feature selection algorithms have omitted the irrelevant features based on mutual information (MI) [20, 21, 22].

The rest of this paper is organized as what follows. In Section 2, we point out the preliminaries and definitions that are used throughout the paper. The proposed method is described in Section 3. In Section 4, the results obtained are demonstrated and analyzed throughout figures and tables. As the final part, Section 6 concludes our research work and suggests some future works.

2. Preliminaries

2.1. PSO

PSO consists of a population of particles in which each particle is a potential solution. After a random initialization of the population, each particle searches through the multi-dimensional search space with a special velocity, and updates its velocity and position based on two factors, its optimum place up to now and the best optimum of all the population. Suppose that D represents the dimension of a search space, $x_{id}(t)$ represents the position of the *i*'th particle at the *d*'th dimension, and $v_i(t)$ is the velocity of the *i*'th particle. The best previously visited position (up to time t) of the *i*'th particle is represented by x_i^{best} and the global best position of a swarm is denoted by x_g^{best}. Also c_1, c_2, r_1, r_2 are fixed random numbers for learning the process. Then the particle's velocity is updated as follows:

$$v_{id}(t+1) = v_{id}(t) + c_1 r_1 \left(x_i^{best} - x_{id}(t)\right) + c_2 r_2 \left(x_g^{best} - x_{id}(t)\right) \qquad (1)$$

and the position of each object is updated by:

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \qquad (2)$$

In the proposed method, each position is a vector with the size of the number of features, in which the presence or absence of feature F_i is represented by 1 or 0 in the i'th element of vector, respectively. The changes in particle velocity can be interpreted as changes in the probability of finding the particle in one state [23].

2.2. Mutual Information

As mentioned earlier, we need a tool to find out the relevancy between the features, and then we can recognize the irrelevant features in order to omit them. In order to understand how much one random variable knows about the mutual information as a benchmark, it is defined as follows:

$$I(X,Y)=H(X)-H(X|Y) \qquad (3)$$

By increasing the mutual information between two variables, the uncertainty between them decreases. Hence, a zero mutual information between two random variables shows their independency. This technique will help us to find out which features are not enough informative to be selected because of the other previously selected features that have almost the same information.

2.3. K-means

One of the most famous algorithms in clustering and classification is K-means. This algorithm selects K points randomly as the initial centers. Then in an iteratively loop, for each object finds the closest center and assigns the object to that class. After all objects are assigned to an appropriate class, the center of each class is recalculated. This process is repeated until the centers converge. The K-means algorithm is shown in Algorithm1

Algorithm 1. K-means.

```
Input: N objects{$x_1, x_2, \dots, x_n$ } to be clustered, k: the number of clusters.
Output: k clusters
    •   Randomly select k objects as initial cluster centers $(c_1, c_2, \dots, c_k)$.
    •   Repeat until centers converge:
        o   For each object $x_i$:
            ▪   Calculate the distance of $x_i$ and each cluster center: $d(x_i, c_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - c_{jk})}$
                # $d(x_i, c_j)$ is the distance between $x_i$ and$c_j$ and $m$ is the dimension of data.
            ▪   Assign each object to the closest cluster.
        o   For each cluster l:
            ▪   Compute the mean of objects in cluster l as the new cluster centers as: $c_l = \frac{1}{N_l}\sum_{i=1}^{N} x_i \times I_{i,l}$
                # $N_l$ is the number of objects in cluster $l$ and $I_{i,l} = \begin{cases} 1; & If\ x_i\ belongs\ to\ cluster\ l \\ 0; & Otherwise \end{cases}$
            ▪   Set $c_l$ as the new center of cluster.
```

3. Proposed Method

The proposed method consists of three main phases: initialization phase, iteration phase, and finding and evaluating the best subset of features. In the rest of this section, we will explain each phase in detail, and in Section 3.4, the final algorithm is presented.

3.1. Initialization Phase

As a pre-process operation, the dataset is divided into two subsets of objects randomly, one as the train set with 80% of objects and the other as the test set with the remaining 20% of objects. This operation is implemented 5 times, and each time the algorithm will learn with train set and then its accuracy will be evaluated by the test set. The average of these five iterations will be reported as the final accuracy.

The mutual information between all pairs of features will be calculated by (3) and saved in a matrix named MI. Thus $MI[i,j]$ represents the mutual information between two vectors F_i and $F_j$such that F_i is the i'th column of the dataset.

NC_{max} is set as the number of the desired particles. Then for each particle p, X_p is initialized randomly. x_p is a binary vector with size of the number of all features, in which the value of the i'th element shows the presence or absence of F_i in x_p. For example, suppose that there are 5 features in the main dataset if particle 1 has a vector as:

[1 0 0 1 1] Then: $X_i = 10011$ which means that this particle chooses F_1,F_4 and F_5 as the selected features. As mentioned earlier, each one of these vectors is a potential solution that should be updated and modified gradually. Each particle also has a velocity vector with the same size of X, which is initialized with a random set. This vector is the base of movement of X in each direction.

3.2. Iteration Phase

This phase, as the heart of our algorithm, consists of three steps that are implemented for each particle. This iteration phase is iterated $N_{\text{itertimes}}$. The steps of this phase are as follow:

Step 1. Updating step: The vector V_i is updated as (2). Then if V is greater than a threshold,X_i is set to 1; otherwise, it is set to 0.

Step 2. Omitting redundant or irrelevant features: The sum of the mutual information between all the features present in X_i is calculated, and if it is greater than the average of matrix *MI*, a redundant feature should be omitted and another feature with minimum similarity to the other selected features should be added instead.

Step 3. Calculating the fitness function: The feature subset that is selected by each particle should be evaluated, and based on the result obtained, the local best and global best should be updated. In order to evaluate this subset, we use the 5-fold technique on the train set. The train set is divided into two subsets, 80% of objects as the sub-train and the other 20% as the sub-test. For 5 times, each time the K-means is implemented on the objects of the sub-train with only the selected features in X_i, and after this learning phase, it is implemented on the sub-test. The mean of this 5 times is regarded as the fitness of the corresponding particle $F(X_i)$. This fitness is compared with the fitness of the local best x_l^{best} and global best, x_g^{best} .The local best is the state of X_i with maximum fitness up to now. The global best is the state with maximum fitness among all features.

3.3. Finding and Evaluating the Best Subset

After performing the 3 previous steps,N_{iter} times, the global best is reported as the selected features. The selected features are trained with the train set and evaluated with the test set in the 5-fold method.

At this stage, the accuracy of the classified test set is computed, and the mean accuracy will be calculated as the final output. The proposed algorithm is presented in Algorithm2.

Algorithm 2. The proposed algorithm.

```
Input: The matrix of data
Output: The subset of selected features
  • Repeat 5 times:
      o Divide data into 80% train and 20% test.
          ▪ Compute matrix MI by computed the MI between all pairs of features $F_m$ and $F_n$ by (3)
          ▪ Set NCmax , Niter, δ
          ▪ For each i in 1: NCmax:
              • Randomly set the binary vector Xi
              • Randomly set the binary vector Vi
          ▪ For each t in 1: $N_{iter}$:
              • For each i in 1: $NC_{max}$:
                  o Update all dimensions of vectors Xi, Vi
                  o Compute $S = \sum_m \sum_n MI[m,n].X_{im}.X_{in}$
                  o If S> δ: find the m and n witch $F_m$ and $F_n$ is maximum.
                  o Substitute one of them with a $F_p$ of minimum mutual information with the other one.
                  o #Now we use 5-fold technique to evaluate each vector X to update local and global
                    optimum.
                  o #$x_i^{best}$is the best of vector Xi and $x_g^{best}$ is the best of all vectors.
  • For 5 times:
      o Divide "train data" into 80% sub-train and 20% sub-test.
      o Implement K-means on sub-train with the features selected by $X_i$ .
      o Implement K-means on sub-test with the features selected by $X_i$ and compute result.
      o Compute Fitness(Xi) as the average result of these five runs.
      o Update $x_i^{best}$ and $x_g^{best}$.
  • Return $x_g^{best}$as the final selected features.
```

4. Experimental Results

We evaluated the performance of the proposed method on 3 classification benchmark datasets: Glass, Vowel, and Wine, given in table 1.

Table 1. Characteristics of UCI datasets used for evaluating the proposed method.

Name	Num_{features}	Num_{objects}	Num_{classes}
Vowel	10	528	11
Glass	9	214	6
Wine	13	178	3

This algorithm was implemented on MATLAB 2008 on a cori7 system with 8G RAM. We run our algorithm 20 times ($N_{iter} = 20$) and compared the results of our proposed method with three meta-heuristic methods, ACO, GA, and PSO feature [25] selection methods, and also with the case without feature selection. This comparison is demonstrated in table 2. In the Vowel dataset, our proposed method gained an average accuracy of 91%, which was much more than the other three algorithms and the case without feature selection. In the Glass dataset, the average accuracy was 98%, which was better than the best of the others and the case without feature selection. And finally, in Wine, the average accuracy, 93%, was very close to the best case.

Table 2. Average accuracy of the proposed method in comparison with 3 other feature selection methods and the case without feature selection.

Name	Proposed method	ACO	GA	PSO	Total features
Vowel	91%	70%	64%	70%	71%
Glass	98%	92%	92%	94%	96%
Wine	93%	74%	84%	95%	97%

5. Conclusion

Nowadays, the progress of information techniques leads to obtain high-dimensional datasets with many different features. This phenomenon, which is called curse of dimensions, can cause some challenges like intractable complexity or misleading information. The feature selection techniques are designed to reduce the dimensionality of the datasets by selecting the most informative features without losing important information for the classification task. It omits the irrelevant and redundant features. In this paper, we proposed a new method based on PSO and mutual information for feature selection. PSO, as a heuristic algorithm, can reduce the complexity and obtain a near-optimal solution. Mutual information can help us to distinguish the relationship between the features and choose the most informative of them.

The algorithm was implemented on three datasets: Vowel, Wine, and Glass. The results obtained were compared with 3 meta-heuristic methods, ACO, GA, and PSO, and also with the case without feature selection. The results obtained show its superiority over them.

References

[1] B. Tang, S. Kay and H. He, "Toward optimal feature selection in naive Bayes for text categorization, " *IEEE transactions on knowledge and data engineering,* vol. 28, pp. 2508-2521, 2016.

[2] K. Yurtkan and H. Demirel, "Feature selection for improved 3D facial expression recognition, " *Pattern Recognition Letters,* vol. 38, pp. 26-33, 2014.

[3] S. Tabakhi, A. Najafi, R. RAnjbar and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization, " *Neurocomputing,* vol. 168, pp. 1024-1036, 2015.

[4] M. Salehi, J. Razmara and Sh. Lotfi, "Development of an Ensemble Multi-stage Machine for Prediction of Breast Cancer Survivability, " *Journal of AI and Data Mining,* vol. 8, pp. 371-378, 2020.

[5] S. Beigi and M. R. Amin Naseri, "Credit Card Fraud Detection using Data mining and Statistical Methods, " *Journal of AI and Data Mining,* vol. 2,pp. 149-160, 2020.

[6] Mirzadeh, Nader and Ricci, Francesco and Bansal, Mukesh, "Feature selection methods for conversational recommender systems, " in *2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*, 2005, pp. 157-165.

[7] Y.S. Jeong, K. S. Shin, and M. K. Jeong, "An evolutionary algorithm with the partial sequential forward floating search mutation for large scale feature selection problems, "*Journal of The Operational research society* vol. 66, pp. 529-538, 2014.

[8] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design, " *IEEE Trans. Syst,* vol. 36, pp. 106-117, 2006.

[9] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems, "*European Journal of Operational Research,* vol. 206, pp. 528-539, 2010.

[10] B. Chen, L. Chen, Ling and Y. Chen, "Efficient ant colony optimization for image feature selection, " *Signal processing,* vol. 93, pp. 1566-1576, 2013.

[11] Z. Zhu, Y. S. Ong and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection, " *Pattern Recognition,*vol. 40, pp. 3236-3248, 2007.

[12] M. Marinaki and Y. Marinakis, Yannis, "A bumble bees mating optimization algorithm for the feature selection problem, " *International Journal of*

Machine Learning and Cybernetics, vol. 7, pp. 519-538, 2016.

[13] H. Yu, G. Gu, Guochang , H Liu, J. Shen and J. Zhao, "A modified ant colony optimization algorithm for tumor marker gene selection, " *Genomics, proteomics and bioinformatics*,vol. 7, pp. 200-208, 2009.

[14] S. W. Lin,K. Ying, Sh. Chen, and Z. Lee, , "Particle swarm optimization for parameter determination and feature selection of support vector machines, " *Expert systems with applications*, vol. 35, pp. 1817-1824, 2008.

[15] S. M. Vieira, L. F. Mendon, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients, " *Applied Soft Computing*,vol. 13, pp. 3494-3504, 2013.

[16] C. L. Huang and J. F. Dun, "A distributed PSO-SVM hybrid system with feature selection and parameter optimization, "*Applied Soft Computing*, vol. 8, pp. 1381-1391, 2008.

[17] B. Xue, L. Cervante, L. Shang, Lin and W. Browne, "A multi-objective particle swarm optimisation for filter-based feature selection in classification problems, " *Connection Science*, vol. 24, pp. 91-116, 2012.

[18] L. Cervante, B. Xue, L. Shang and M. Zhang, "A multi-objective feature selection approach based on binary pso and rough set theory, " *European Conference on Evolutionary Computation in Combinatorial Optimization*, 2013, pp. 25-36.

[19] C. E. Shanon , "A mathematical theory of communication, "*Bell system technical journal*, vol. 27,pp. 379-423, 1948.

[20] H. Peng, F. Long and Ch. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, "*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226-1238, 2005.

[21] M. Rahmaninia, P. Moradi, "Osfsmi: Online stream feature selection method based on mutual information, " *Applied Soft Computing*, vol. 68, pp. 733-746, 2018.

[22] N. Bi, J. Tan, J. H. Lai and Ch. Suen, "High-dimensional supervised feature selection via optimized kernel mutual information, " *Expert Systems with Applications*,vol. 108, pp. 81-95, 2018.

[23] J. Kennedy, R.C. Eberhart, "A discrete binary version of the particle swarm algorithm, In: Systems, Man, and Cybernetics, " in *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*, 1997, pp. 4104-4108.

انتخاب ویژگی با استفاده از بهینه‌سازی ازدحام ذرات و اطلاعات متقابل

زهرا شجاعی[۱]، سیدابوالفضل شاهزاده فاضلی[۲،*]، الهام عباسی[۳] و فضل الله ادیب نیا[۴]

[۱و۲و۳] گروه علوم کامپیوتر، دانشگاه یزد، یزد، ایران.

[۴] دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران.

ارسال ۲۰۱۹/۱۲/۰۸؛ بازنگری ۲۰۲۰/۰۲/۰۵؛ پذیرش ۲۰۲۰/۰۷/۰۶

چکیده:

امروزه انتخاب ویژگی، بعنوان روشی برای بهبود کارایی متدهای طبقه‌بندی، مورد توجه بسیاری ازعلاقمندان حوزه‌ی یادگیری ماشین واقع شده است. از آنجایی‌که در پردازش و انجام عملیات برروی یک ماتریس، ابعاد یک ماتریس نقش مهمی را ایفا میکند، کاهش تعداد ویژگی‌ها با انتخاب بهترین زیرمجموعه‌ی ویژگی‌ها تاثیر بسزایی در کارایی الگوریتم‌ها خواهد داشت. از سویی یافتن این بهترین زیرمجموعه با استفاده از مقایسه‌ی تمامی زیرمجموعه‌های ممکن با یکدیگر، حتی هنگامی‌که اندازه مجموعه کوچک باشد نیز کاری طاقت‌فرسا می‌باشد. از این‌رو باکارهای تحقیقاتی زیادی تلاش کرده‌اند که با استفاده از روش‌های فراابتکاری راه‌حلی نزدیک به بهینه برای این مساله بیابند. دراین مقاله ما روشی جدید برای انتخاب ویژگی معرفی کرده‌ایم که به انتخاب ویژگی‌هایی با بیشترین اطلاعات می‌پردازد و ویژگی‌هایی که کمتر اطلاعات جدیدی برای ارائه دارند، یا حتی حاوی اطلاعات گمراه‌کننده یا نامرتبطی هستند را از زیرمجموعه حذف می‌کند. دراین روش پیشنهادی برای یافتن زیرمجموعه‌ای از ویژگی ها که نزدیک به بهینه باشند، به‌سراغ روش بهینه‌سازی ازدحام ذرات رفته‌ایم. همچنین بمنظور شناسایی ویژگی‌های نامرتبط یا اضافی نیاز است که ارتباط میان ویژگی های مختلف کشف شود. از میان توابع گوناگون همبستگی، از روش اطلاعات متقابل استفاده کرده‌ایم. درپایان برای محک و ارزیابی روش پیشنهادی، الگوریتم را برروی پایگاه‌داده های Glass, Vowel و Wine پیاده‌سازی کرده‌ایم. مقایسه‌ی نتایج بدست آمده با چهار روش شناخته شده‌ی امروزی، بیانگر قدرت و کارایی بالای روش پیشنهادی می‌باشد.

کلمات کلیدی: انتخاب ویژگی، اطلاعات متقابل، بهینه سازی ازدحام ذرات، طبقه‌بندی، یادگیری ماشین.