ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 141

International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017)

Feature Selection on the Basis of Rough Set Theory and Univariate Marginal Distribution Algorithm

Bin Wei[1,*], Minqing Zhang[1], Longfei Liu[1] and Jing Zhao[2]

[1]Key Laboratory of Network & Information Security of APF, Engineering College of APF, Xi'an, China

[2]College of Science Xijing University, Xi'an, China

***Abstract*—Feature selection is an important preprocessing step in machine learning. The aim of feature selection is to find an optimal subset from original features that satisfies a criterion. Rough set theory (RST) is one of the most effective ways to solve feature selection problem, but RST is inefficient in large scale datasets. In order to solve this problem, in this paper, we proposed a novel feature selection algorithm RSUMDA on the basis of univariate marginal distribution algorithm. RST was used to obtain the significance of each feature as the original probability of UMDA and then UMDA was to search the optimal feature subset that using the number of the selected feature and the accuracy of the classifier as fitness function. Experimentation was carried out in 4 UCI datasets. The results showed that our algorithm could effectively reduce the number of the features, improve the accuracy of the classifier and quicken the convergence rate.**

Keywords-feature selection; rough set theory; UMDA

I. INTRODUCTION

With the rapid development of computer network and sensor technology, we can measure more and more features. But in such huge features may contain a lot of redundant, irrelevant and noise ones. Consequently much research has been performed on choosing useful information from such huge features[1]. Feature selection is a useful preprocessing step for removing irrelevant and noise data, reducing dimensionality, enhancing output comprehensibility and improving learning accuracy. Feature selection is widely used in the field of data mining, machine learning and pattern recognition[2]. The aim of feature selection is to choose an optimal features subset while retaining as much as original information.

Rough Set Theory (RST) has been proposed by Pawlak[3] to deal with vague, imprecise and uncertain information. RST is one of the most effective approaches to solve feature selection problem. Generally, using RST to solve feature selection problem can be divided into two kinds: hill-climbing methods and stochastic methods [4].

The hill-climbing methods usually use rough set feature significance as heuristic information. Some of the hill-climbing methods start with an empty set or a core set and then add a feature in turn according to the important of the feature from candidate set. The others begin with full feature set and successively eliminate an irrelevant feature[5]. However, hill-climbing methods often lead to non-minimal feature combination. Therefore, many researchers use stochastic methods for RST feature selection. Multi-objective genetic algorithm was applied for RST feature selection method to the face recognition problem [6]. A RST and genetic algorithm (GA)-based feature selection method was proposed in literature [7]. Literature [8] used PSO-based RST feature selection method to search the optimal subset. A multi-objective ant colony optimization was proposed for rough feature selection [9]. In general, stochastic algorithms can obtain strong robustness at the expense of increasing computation time.

Although traditional Evolution Algorithms (EAs), such as GA, PSO and ACO, are an effective way to solve feature selection problem[10], traditional EAs have some drawbacks: firstly, many parameters need to be tuned; secondly, they are easy to fall in local optimal solution; thirdly, computation time accumulates exponentially as population size. In order to overcome these shortcomings, a new branch of EAs, namely Estimation of Distribution Algorithms (EDAs)[11], was employed in this paper to solve the feature selection problem. Univariate Marginal Distribution Algorithm (UMDA) is one of the EDAs, which is the combination of statistics learning theory and EAs. Therefore, a two stages algorithm named RSUMDA was proposed in this paper. In the first stage, RST was used to calculate the significance of each feature which was used as the original probability of UMDA. In the second stage, the optimal subset was selected by UMDA. RSUMDA adopted the accuracy of the classifier and the number of the features as heuristic information. Compared with other stochastic methods, our proposed method not only decreased computing time and the amount of attributes, but also improved the performance of classifier.

II. OVERVIEW OF TECHNIQUES

A. Rough Set Theory

RST was proposed by Pawlak to deal with imprecise, vague and uncertain data[12]. It uses the repository that we know to approximate the imprecise or uncertain data and discover the hidden knowledge. The main advantage of RST is that it doesn't need for providing any additional or prior information of the dataset, then avoiding the subjectivity. Feature selection is the main application of RST. In this selection we present the basic concepts of RST.

B. UMDA

EDAs, which is the new branch of EAs, is a relatively novel heuristic searching algorithm. The difference between EDAs and the other evolutionary search algorithms is that the

evolution strategy they used from one generation to the next. EDAs use a candidate solutions' spatial distribution probability model to replace conventional evolutionary operators such as crossover and mutation. UMDA[13] was proposed by German scholar Mühlenbein, which is one of EDAs and assumes the variable independent of each other. It can effectively solve high dimension problem. UMDA has some advantages over traditional EAs: fewer parameters need to be tuned; an elitist strategy has been adopted and not easy to fall into local optimal. This algorithm only needs to pre-set the population size, the maximum number of iterations and the number of the optimal samples.

III. RSUMDA-BASED FEATURE SELECTION METHOD

In this section we have introduced our proposed method. We employed RST to calculate the significance of each feature, and then used the feature significance as initial probability to quicken the convergence speed of the UMDA method. The main process of our method contained three parts:

1) Discretize the dataset.

2) Measure the significant of each attribute as original probability of UMDA by RST.

3) Apply the UMDA-based heuristic strategy to obtain optimal feature subset.

A. Discretization

Discrete step was necessary because RST can only deal with discrete value. In this paper we used Naïve Scaler Algorithm (NSA) to discretize the data, which is a well-known discretization algorithm. It didn't need to set any parameters that only used the condition and decision attributes to discretization. The main steps of the NSA are show in algorithm 1.

Algorithm 1
Input: continuous decision table $S=<U,C\cup\{d\},v,f>$
Output: discrete decision table $S^p=<U,C\cup\{d\},v^p,f^p>$
Step 1 breakpoints_set=Ø
Step 2 Arrange each continuous attributes $a\in C$ from small to large to get a sequence of samples $x_1, x_2,\ldots, x_N$(N is the number of the samples), and $f_a(x_1)\le f_a(x_2)\le\ldots\le f_a(x_N)$, $f_a(u_i)$ denotes the condition attribute a's value of sample x_i
Step 3 Scan the continuous attribute values in turn. If the corresponding decision attribute value is different, add the average of condition attribute value to the breakpoints_set
For i=1 to (N-1)
If $f_a(u_i)\neq f_a(u_{i+1})$
Breakpoints_set= Breakpoints_set$\cup((f_a(u_i)+f_a(u_{i+1}))/2)$
end if
end for
Step 4 Use the breakpoints from the Breakpoints_set to discretize the dataset
Step 5 Output discrete decision table $S^p=<U,C\cup\{d\},v^p,f^p>$

B. RSUMDA-based feature selection method

RSUMDA was a two stages method. First, the significance of each attribute was calculated by RST formula. Then the importance of each attribute was used as original probability model of UMDA. The same as other EAs, individual encoding and the fitness function designing are the most important problem in RSUMDA. In this paper, the dimension of the individual was equal to the sum numbers of all features. The detailed of the individual description was described as follows:

$$a_1, a_2, \ldots a_M$$

where M meant the number of the condition attributes, $a_i \in \{0,1\}$, namely, if a_i equaled to 1 denoted the attribute was selected, otherwise was not. Fig.1 shows examples of encoding for RSUMDA. The dataset contains a total of ten condition attributes. $C=\{a_1, a_2,\ldots a_{10}\}$ is the set of condition attributes. The selected subset is $C_s=\{a_1, a_2, a_6, a_8, a_9, a_{10}\}$

All the features

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
					↓				
1	1	0	0	0	1	0	1	1	1

FIGURE I. EXAMPLES OF ENCODING FOR RSUMDA

The fitness function was used to guide the population evolution direction. The target of feature selection was to choose the optimal subset was with least length and highest classification quality. Thus, we combined the two criteria into an objective fitness function that looked for the optimal equalization between the features' number and the accuracy. As mentioned above the fitness function was designed as follows:

$$Fit = \alpha * f_1 + (1-\alpha) * f_2 \quad (1)$$

$$f_1 = (M-q)/M$$

where M was the total number of condition attributes, q was the number of attributes that selected by RSUMDA. f_2 was the accuracy of the classifier. f_1 was negatively correlated with the number of the selected attributes. $\alpha \in (0,1)$ was the weight coefficient between the attribute's number and the accuracy. Maximize fitness function was the goal of RSUMDA. The main steps of RSUMDA-based feature selection algorithm are shown in algorithm 2.

Algorithm 2
1 Input: original decision table $S=<U,C\cup\{d\},v,f>$ and discrete decision table $S^p=<U,C\cup\{d\},v^p,f^p>$
2 The significance of each condition attribute was calculated by S^p and formula (9) as original probability of UMDA
3 Set pop size to W and the maximum number of iterations to T
4 Generate the initial population according to original probability
5 Calculate individuals' fitness function according to the formula (11), S uses to calculate the accuracy of the classifier
6 Select the best $F=W/2$ individuals as optimal population D^{sel}_{t-1}, set $t=t+1$
7 Calculate the joint probability distribution $P_t(x)$ according to formula (10)
8 Generate $W-F$ new individuals by sampling from $P_t(x)$
9 If $t\leqslant T$ or when t>5, the maximal fitness function in this iteration Fit(t)≠Fit(t-5) , go to 5, else, stop
9 Output the individual with the maximal fitness value

IV. EXPERIMENTS AND RESULTS

Our experiments were conducted on the 4 datasets from U.C. Irvine Machine Learning Repository [14]. We removed the instances from the datasets if they had missing values. The

details of the datasets are listed in Table 1.

TABLE I. DESCRIPTION OF DATASETS

Datasets	Condition attributes	Samples	class
splice	60	3190	3
Isolet	617	1200	4
Libra	90	360	15
Musk	166	476	2

Support vector machine (SVM) is one of the widely used classifier. Thus we adopted SVM with radial basis kernel function (RBF) as classifier to obtain the accuracy[15]. Moreover, k-fold cross-validation approach was employed in the experiments, with k=10.

We compared the result of our method with that of UMDA and the method used in [16]. Furthermore, the influence of the different fitness functions on the results was also tested in this paper. All the algorithms of the population size and maximum iteration were set to 50 and 300 respectively. In the following tables, the superscript expressed different fitness function. If the superscript was 1 meant that only used accuracy as fitness function. The superscript was set to 2 showed that used the fitness function the same as our proposed method (formula 11). Methods used 1 as fitness function was presented as Z^1. Z^2 was methods used formula 11 as fitness function. Thereinto, Z was PSO, UMDA or RSUMDA.

Table 2 displays the influence of α in the splice dataset. The accuracy decreased gradually with the increase of α. When α was large, the selected number of features played a dominance role in fitness function, namely the less of the number the higher of the fitness function. The selected number of features and the accuracy could achieve a balance when α was 0.2. Therefore in the following tests, the weight coefficient (α) was set as 0.2.

TABLE II. INFLUENCE OF THE WEIGHT COEFFICIENT A

splice dataset	PSO^2		$UMDA^2$		$RSUMDA^2$	
	Accuracy	Num	Accuracy	Num	Accuracy	Num
α =0.1	85.06	29	88.31	12	87.93	11
α =0.2	85.46	22	90.65	7	90.84	6
α =0.3	83.37	20	89.68	6	90.34	6
α =0.4	83.00	20	89.40	6	90.34	5
α =0.5	83.09	19	90.34	5	89.21	5
α =0.6	80.53	20	88.66	5	90.00	3
α =0.7	80.00	20	84.65	4	88.59	4
α =0.8	75.68	19	76.09	3	81.50	3
α =0.9	66.93	15	63.28	1	62.37	1

The results in terms of accuracy and feature's number were summarized in Table 3. In Table 3, the second column showed the accuracy was obtained by using all the condition attributes. From third to eighth column presented the accuracy that was acquired by the subset of the features. PSO^1-based method had higher precision than $UMDA^1$-based method in Musk datasets. PSO^1-based method had higher accuracy than $RSUMDA^1$-based method in Musk datasets. Therefore PSO had lower accuracy than other methods in the most of the datasets whatever which fitness function was adopted.

TABLE III. PREDICTIVE ACCURACY OF SVM CLASSIFIER WITH RBF KERNEL FOR THE METHODS

Datasets	PSO^1	PSO^2	$UMDA^1$	$UMDA^2$	$RSUMDA^1$	$RSUMDA^2$
splice	84.43	85.46	89.95	90.65	88.87	90.84
Isolet	98.75	99.08	99.33	99.33	99.58	99.16
Libra	60.33	61.00	67.33	67.33	68.00	66.33
Musk	66.74	67.20	66.51	66.74	66.51	66.97

TABLE IV. THE CHANGE OF THE ACCURACY AND NUMBER BETWEEN DIFFERENT FITNESS FUNCTION FOR SVM CLASSIFIER

Datasets	PSO^1- PSO^2		$UMDA^1$- $UMDA^2$		$RSUMDA^1$- $RSUMDA^2$	
	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)	Num
splice	-1.0313	9	-0.7000	27	-1.9688	13
Isolet	-0.3333	46	0	240	0.4166	138
Libra	-0.6667	11	0	26	1.6667	12
Musk	-0.4651	12	-0.2326	39	-0.4651	6

TABLE V. THE REDUCTION OF THE ACCURACY AND NUMBER AMONG VARYING METHODS FOR SVM CLASSIFIER WITH RBF KERNEL

Datasets	PSO1- RSUMDA1		UMDA1- RSUMDA1		PSO2- RSUMDA2		UMDA2- RSUMDA2	
	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)	Num
splice	-4.4375	12	1.0813	15	-5.3750	16	-0.1875	1
Isolet	-0.8333	134	-0.2500	149	-0.0834	226	0.1666	47
Libra	-7.6667	13	-0.6667	16	-5.3333	14	1.0000	2
Musk	0.2326	28	0	36	0.2326	22	-0.2325	3

TABLE VI. WHICH GENERATION OF UMDA CONVERGED BY USING ACCURACY AND FORMULA 1 AS FITNESS FUNCTION FOR SVM.

Datasets	UMDA1		RSUMDA1	
	accuracy	formula 1	accuracy	formula 1
splice	23	19	19	15
Isolet	52	46	45	44
Libra	65	51	48	46
Musk	221	217	84	74

Table 4 presents the change of accuracy and number between different fitness function. In splice, Isolet, Libra and Musk datasets PSO^2-based method had better accuracy than PSO^1-based method. In splice and Musk $UMDA^2$-based method had higher precision than $UMDA^1$-based method and the two methods had the same accuracy in Isolet and Libra datasets. In splice and Musk $RSUMDA^2$-based method was more precise than $RSUMDA^1$-based method. Though the accuracy was reduced a little when used formula 1 as fitness function in the most of the datasets, the number of the features were decreased in varying degrees. Therefore formula 1 as criterion for selecting optimal subset looked for a balance between accuracy and number.

Table 5 displays the change of accuracy and number between our proposed method and others. Minus in accuracy column and number column shows the algorithm behind the minus has higher precision and more numbers of features than algorithm before the minus respectively. PSO^1-based method had higher accuracy than $RSUMDA^1$-based method in Musk dataset. $UMDA^1$-based method had better accuracy in 2 datasets and less number of the features in dermatology dataset than $RSUMDA^1$-based method. Compared with PSO^2-based method our proposed method had improved the accuracy and reduced the number of features a lot simultaneously. Only in Musk dataset the accuracy of PSO^2-based method was better than $RSUMDA^2$-based method. In short, the use of RST enhanced the performance of the classifier and reduced the number of the features.

Table 6 displays in which generation of UMDA and RSUMDA converged by using accuracy and formula 1 as fitness function. RSUMDA could converge quickly compared with UMDA. Therefore, the use of RST accelerated the convergence rate in the most of the datasets no matter which fitness function was chosen.

We could summarize the main point by experiments as follows: firstly, the accuracy of UMDA-based method was higher than PSO-based method in the most of the datasets; secondly, the number of the features dropped by using formula 1 as fitness function even though the accuracy decreased slightly in some datasets; thirdly, the convergence rate of UMDA was enhanced by RST. Hence our proposed method could effectively solve feature selection problem.

V. CONCLUSIONS

In this paper, we have proposed a RSUMDA to solve feature selection problem. Our method was the two stage processes that combined of RST filter and UMDA wrapper methods. Firstly, RST was used to calculate the significant of each attribute as original probability for generating the UMDA's original population. Secondly, UMDA was employed to obtain the optimal subset that gained balance between number of the features and the accuracy of the classifier. Moreover, our algorithm didn't need to tune many parameters and was easy to realize. The experimental results showed that our method as preprocessing steps of the classifier could effectively reduce the dimension, quicken the convergent speed and improve accuracy on the most of the datasets we used in this paper.

ACKNOWLEDGMENT

This study was supported by the Scientific research program funded by Shaanxi provincial education department (program NO. 15JK2187), Scientific research program funded by Xijing University (program NO. XJ160235), National social science foundation (program NO. 16BTJ033), Foundation of engineering college of APF (WJY201518).

REFERENCES

[1] Katrutsa, A.S., Vadim, Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. EXPERT SYSTEMS WITH APPLICATIONS, 2017. 76 (15): p. 1-11.

[2] Li, Y.Y., Yuantao; Li, Guoyan, A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection. MECHANICAL SYSTEMS AND SIGNAL PROCESSING 2017(91): p. 295-312.

[3] Dai, J.H., Huifeng; Zhang, Xiaohong, Catoptrical rough set model on two universes using granule-based definition and its variable precision extensions. INFORMATION SCIENCES 2017. 390 p. 70-81.

[4] Wang, X.Y., et al., Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters, 2007. 28(4): p. 459-471.

[5] Ge, X.W., Pei; Yun, Ziqiu, The rough membership functions on four types of covering-based rough sets and their applications. INFORMATION SCIENCES, 2017. 390 p. 1-14.

[6] Mazumdar, D. and S. Mitra, Evolutionary-rough feature selection for face recognition. Transactions on Rough Sets. XII2010: Springer. 117-142.

[7] Xin, P. and Z. Suli, Ensemble remote sensing classifier based on rough set theory and genetic algorithm. 2010 18th International Conference on Geoinformatics, 2010.

[8] Wang, X.Y., et al., Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma. Computer Methods and Programs in Biomedicine, 2006. 83(2): p. 147-156.

[9] Liangjun, K., et al., A multiobjective ACO algorithm for rough feature selection. 2010 Second Pacific-Asia Conference on Circuits,Communications and System (PACCS 2010), 2010: p. 207-210210.

[10] Alba, E.M., J.; Dorronsoro, B, Theory and practice of cellular UMDA for discrete optimization. PARALLEL PROBLEM SOLVING FROM NATURE - PPSN IX, PROCEEDINGS 2006. 4193 p. 242-251.

[11] Gao, S.d.S., Clarence W, A modified estimation distribution algorithm based on extreme elitism. BIOSYSTEMS 2016. 150 p. 149-166.

[12] Pacheco, F.C., Mariela; Sanchez, Rene-Vinicio, Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery. EXPERT SYSTEMS WITH APPLICATIONS, 2017. 71 p. 69-86.

[13] Xu, Z.W., Yirui; Li, Sheng, Immune Algorithm Combined with Estimation of Distribution for Traveling Salesman Problem. IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING, 2016. 11 (1): p. 142-154

[14] Blake, C.L. and C.J. Merz, UCI Repository of machine learning databases. Department of Information and Computer Science,University of California, Irvine,CA,, 1998. http://www.ics.uni.edu/~mlearn/MLRepository.htm.

[15] Bin, W., A hybird algorithm for disease association stduy. Journal of biomedical science and engineering, 2016. 9(10): p. 129-136.

[16] Lin, S.W. and S.C. Chen, PSOLDA: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis. Applied Soft Computing, 2009. 9(3): p. 1008-1015.