An Application of PSO Algorithm and Decision Tree for Medical Problem

Meng-Chang Tsai, Kun-Huang Chen, Chao-Ton Su, and Hung-Chun Lin

***Abstract*—**In this study, we propose a novel method for medical problem, it is the integration of particle swarm optimization (PSO) and decision tree (C4.5) named PSO + C4.5 algorithm. To evaluate the effectiveness of PSO + C4.5 algorithm, it is implemented on 5 different data sets of life sciences obtained from UCI machine learning databases. Moreover, the results of PSO + C4.5 implementation are compared to logistic regression (LR), back propagation neural network (BPNN), support vector machine (SVM), and decision tree (C4.5). The accuracy index shows that PSO + C4.5 algorithm outperforms the other methods.

***Keywords*—**Decision tree, Feature selection, Particle swarm optimization.

I. Introduction

DATA mining is a process of knowledge discovery in databases (KDD) to discover the useful information from the existing database. This data mining is accomplished in different tasks; classification is a popular data mining task. In medical research, classification task is widely applied to medical data to aid physicians to determine the prevalence of diseases. However, it is challenging for classification in medical filed due to data are large, complex, heterogeneous, hierarchical, and high-dimensional.

In high dimensional classification, it can be problem because a higher number of features used do not always imply the higher performance for classification. Conversely, the classifier performance can even decrease when those features are irrelevant or correlated. Therefore, feature selection is considered to use in pre-processing in data mining before applying classifiers to a data set. Thus, the good feature selection method leads to the high classification accuracy and reduces computational cost.

Feature selection plays an important role to choose the important features for classification while the classifier performances remain high. Basically, the optimal feature selection is categorized into two types: filter method and wrapper method, these methods are exhaustive search. To select features for filter methods, the features are scored and ranked based on statistical criteria such as t-test, chi-square test, Wilcoxon Mann-Whitney test, mutual information, Pearson correlation coefficients, and principal component analysis [2,6]. Then, the features with the highest score are selected as optimal features. In wrapper method, feature selections are wrapped with classifier algorithms. Thus, the feature subsets are evaluated by learning algorithms, the feature subset with highest prediction accuracy is an optimal subset. Although, filter and wrapper methods try to find the optimal subset of features, these searching strategies are exponentially prohibitive in exhaustive searches; moreover, searching for an optimal feature subset from a high dimensional feature space is known to be an NP-complete problem [5] so the high time consuming is still an unsolved problem. Therefore, feature selection through meta-heuristic methods is interesting for solving the problem in this study.

Genetic algorithm (GA) and particle swarm optimization (PSO) are meta-heuristic methods, the GA algorithm is a general adaptive optimization search methodology based on a direct analogy to Darwinian natural selection and genetics in biological systems. It is widely used to solve the optimization problems which the quality of solution is evaluated by fitness function [17]. The PSO is a popular algorithm, firstly proposed by Kennedy and Eberhart [4]. This algorithm was discovered by simulation the movement of social behavior of bird flocking and fish schooling. Recently, GA and PSO algorithms have been applied and shown high performances in many researches for feature selection problems [5,8,12,13]. Notably, Gheyas and Smith [5] concluded that both GA and PSO approaches had less computational time than exhaustive search. Moreover, Tu et al. [12] and Ziari and Jalilian [18] concluded that PSO approach was more effective but needed fewer parameters and computational time. Consequently, PSO is considered to apply in various research fields.

The PSO is an informative search and played important search for feature selection. Decision tree (C4.5) is a conventional classifier which has been used for classification problem. However, there are very few researchers use the combination of PSO and C4.5 to improve the classification performance in medical problem.

Meng-Chang Tsai is a Deputy R & D division in Institute of Nuclear Energy Research, Taiwan (e-mail: channingtsai@gmail.com).

Kun-Huang Chen is a postdoctoral in School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan (e-mail: khchen@tmu.edu.tw).

Chao-Ton Su is a Chair Professor with the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan (e-mail: ctsu@mx.nthu.edu.tw).

Hung-Chun Lin is a postdoctoral in National Tsing Hua University, Hsinchu, Taiwan (corresponding author to provide phone: +886-3-5715131; fax: +886-3-5722204; e-mail: d9534801@oz.nthu.edu.tw).

II. Methods

A. Particle swarm optimization (PSO)

Kennedy and Eberhart [4] proposed a method for optimization of continuous nonlinear function using particle swarm methodology, which is well known as Particle swarm optimization (PSO). This method was discovered based on animal social behavior: the movement of bird flocking and fish schooling. In PSO algorithm, a swarm consists of *N* particles moving around the D-dimensional space, the *i*th particle is vector $X_i = [x_{i1}, x_{i2}, ..., x_{iD}]$, and the velocity vector is $V_i = [v_{i1}, v_{i2}, ..., v_{iD}]$. The record of position of its previous best performance is $PB_i = [pb_{i1}, pb_{i2}, ..., pb_{iD}]$ and the best performance so far in the neighborhood is $GB_i = [gb_{i1}, gb_{i2}, ..., gb_{iD}]$.

B. Decision tree (C4.5)

C4.5 has been used in real-world problems especially medical decision making because the methods can simultaneously provide high classification accuracy, simple representation of gathered knowledge. It is an improved algorithm of ID3 algorithm, presented by Quinlan [10]. C4.5 is a top-down algorithm and builds a decision tree model using a recursive process (also known as divides and conquer strategy). It uses information gain as splitting criteria to build a decision tree, an attribute with the most information which is computed on training set is first selected , the next one is selected the most informative from the remaining attributes, and so on. C4.5 algorithm can handle numeric attributes and missing values [11].

C. Proposed approach

We integrate PSO algorithm and C4.5 to address the feature selection problem. In this proposed approach, the PSO is employed for important feature selection, and C4.5 method is used as a fitness function of the PSO in order to test for the efficiency of the set of selected features.

III. Experiment and results

A. Environment

In this section, to verify the performance of PSO + C4.5 for feature selection is shown. We test the proposed algorithm on 5 data sets presented in Table I. These data sets are obtained from UCI machine learning databases (http://archive.ics.uci.edu/ml/). All features of these data sets are numeric and do not have missing values.

Five-fold cross-validation is applied for classification results evaluation on data sets (except SPECT heart and SPECTF heart data sets, because they have contained both training and testing sets), so that the bias caused by random sampling for training and testing sets can be reduced [14].

In this study, the proposed algorithm is coded in Visual studio C++ 2005. The parameters of PSO are set follow [3]; hence the cognitive learning factor (c_1) and the social learning factor (c_2) are set at 2 for each, the values of lower bound of velocity (v_{min}) and upper bound of velocity (v_{max}) are -4 and 4, respectively. The number of particles in swarm is the number of features. The inertia weight (w) is set at 0.4. The process is repeated until either the fitness of the given particle was 1.0 or the number of iterations was achieved by default value T. Herein, T for PSO + C4.5 algorithm are set at 100.

TABLE I
Summarization of Data Set Characteristics

	# of features	# of samples	# of classes
Breast Tissue	10	106	6
Haberman's Survival	3	306	2
Liver Disorders	7	345	2
Parkinsons	23	197	2
Pima Indians Diabetes	8	768	2

B. Numerical experiments

To evaluate the effectiveness of the proposed method, logistic regression (LR), back propagation neural network (BPNN), support vector machine (SVM) and decision tree (C4.5) methods are also implemented on the data sets, then the results are compared to PSO + C4.5 algorithm.

Logistic regression (LR) is a statistical method, it has been generally used in various fields and it is well known as the gold standard method in prediction task [1]. The back propagation neural network (BPNN) is the most popular of the neural network applications. Support vector machine is a classification method widely used in different research areas. Decision tree (C4.5) is a well-known classification method because this method can simultaneously provide high classification accuracy, simple representation of gathered knowledge. Though many researches showed that BPNN, SVM have had high accuracy performance , these methods still have drawback is that they have high time consuming [6,7,9,15,16].

The implementation of SVM, BPNN are followed [3], a library for support vector machines (LIBSVMs) is used as a tool for SVM implementation; by using this tool, an efficient parameter selection tool using cross-validation via parallel grid search under the kernel of the radial basis function type is provided for SVM implementation. BPNN is implemented using professional II PLUS software, the parameters of BPNN, which those are the learning rate, momentum, and number of hidden nodes, are optimized by trial and error to find the combinations with the minimum root mean square error. The classification accuracy values are presented in Table II. The results clearly show that PSO + C4.5 algorithm is the best of the other methods. Fig. 1 shows the mean classification accuracy.

TABLE II
THE PERCENTAGE OF AVERAGE CLASSIFICATION ACCURACY OF THE 5 DATA SETS

	BPNN	LR	SVM	C4.5	PSO + C4.5
Breast Tissue	21.60	68.83	51.82	60.35	63.16
Haberman's Survival	73.50	73.50	73.50	70.91	72.53
Liver Disorders	57.97	67.83	58.26	64.93	73.04
Parkinsons	75.38	83.59	86.15	82.56	90.77
Pima Indians Diabetes	65.10	76.65	76.65	76.22	75.52

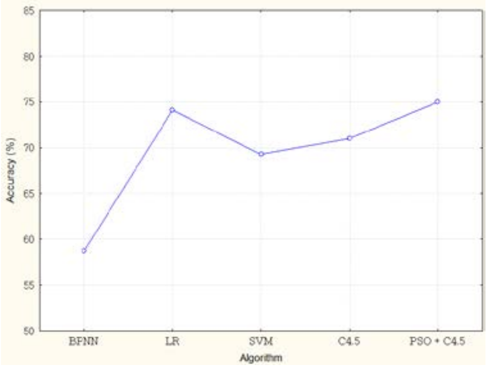

Fig. 3 The mean for classification accuracy

IV. CONCLUSION

In this study, we present a new method for feature selection; it is the integration of particle swarm optimization (PSO) and decision tree (C4.5). This proposed method is applied on 5 data sets obtained from UCI machine learning databases. The accuracy index is used to evaluate the effectiveness of the proposed method; in addition, analysis of variance is used to test the difference classification performance among LR, BPNN, SVM, C4.5 methods, and PSO + C4.5 algorithm. The experimental results show that PSO + C4.5 algorithm outperforms the other methods.

REFERENCES

[1] T. Badriyah, J. S. Briggs, and D.R. Prytherch, “Decision trees for predicting risk of mortality using routinely collected data,” *Int. J. Soc. Hum. Sci.*, vol. 6, pp. 303-306, 2012.

[2] L. F. Chen, C. T. Su, and K. H. Chen, “An improved particle swarm optimization for feature selection,” *Intell. Data Anal.*, vol. 16, no. 2, pp. 167-182, 2012.

[3] L. F. Chen, C. T. Su, K. H. Chen, and P. C. Wang, “Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis,” *Neural Comput. Appl.*, 2011, DOI: 10.1007/s00521-011-0632-4.

[4] J. Kennedy and R. C. Eberhart, “Particle swarm optimization,” in *IEEE Int. Conf. on Neural Networks*, pp. 1942-1948, 1995.

[5] I. A. Gheyas and L. S. Smith, “Feature subset selection in large dimensionality domains,” *Pattern Recogn.*, vol. 43, no. 1, pp. 5-13, 2010.

[6] Y. H. Hung, “A neural network classifier with rough set-based feature selection to classify multiclass IC package products,” *Adv. Eng. Inform.*, vol. 23, no. 3, pp. 348-357, 2009.

[7] C. H. Li and S. C. Park, “Combination of modified BPNN algorithms and an efficient feature selection method for text categorization,” *Inform. Process. Manag.*, vol. 45, no. 3, pp. 329-340, 2009.

[8] Y. Liu, Z. Qin, Z. L. Xu, and X. S. He, “Feature selection with particle swarms,” *Computational and Information Science, Lecture Notes in Computer Science*, vol. 3314, pp. 425-430, 2004.

[9] E. G. Ortiz-García, S. Salcedo-Sanz, A. M. Perez-Bellido, and J. A. Portilla-Figueras, “Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions,” *Neurocomputing*, vol. 72, no. 16-18, pp. 3683-3691, 2009.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1993.

[11] L. Rokach and O. Maimon, *Decision trees, Data Mining and Knowledge Discovery Handbook*, pp. 165-192, 2005.

[12] C. J. Tu, L. Y. Chuang, J. Y. Chang, and C. H. Yang, “Feature selection using PSO-SVM,” *IAENG Int. J. Compu. Sci.*, vol. 33, no. 1, pp. 1-6, 2007.

[13] H. Vafaie and K. D. Jong, “Genetic algorithms as a tool for feature selection in machine learning,” in *4th Int. Conf. on Tools with Artificial Intelligence*, pp. 200-203, 1992.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco, CA: Morgan Kaufmann, 2005.

[15] Z. Yu, M. J. Kim, K. S. Park, and S. H. Kim, “Solid convex-hull sequential support vector machine,” in *Int. Conf. on Broadband, Wireless Computing, Communication and Applications*, pp 181-185, 2010.

[16] H. Zhang, A. C. Berg, M. Maire, and J. Malik, “SVM-KNN: Discriminative nearest neighbor classification for visual category recognition,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2126-2136, 2006.

[17] L. Zhuo, J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian, “A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVII, no. B7, pp. 397-402, 2008.

[18] I. Ziari and A. Jalilian, “A new approach for allocation and sizing of multiple active power-line conditioners," IEEE Trans. Power Deliver., vol. 25, no. 2, pp. 1026-1035, 2010.

Meng-chang Tsai is currently a Deputy R & D division in Institute of Nuclear Energy Research, Taiwan. His research interests include thermal issue, vapor chamber heat spreader, and optimization analysis etc.

Kun-Huang Chen received his PhD degree in Industrial Engineering and Engineering Management from National Tsing Hua University, Taiwan in 2011. He is currently a postdoctoral in School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan. His research interests include data mining and quality engineering.

Chao-Ton Su received the Ph.D. degree in industrial engineering from the University of Missouri, Columbia. He is currently a Chair Professor with the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan. His research activities include quality engineering and management, operation management, and data mining and its applications. Prof. Su is a member of the Institute of Industrial Engineers, Chinese Institute of Industrial Engineers, and Chinese Society for Quality. He is a senior member of the American Society for Quality and an Academician of the International Academy for Quality.

Hung-Chun Lin received his PhD degree in Industrial Engineering and Engineering Management from National Tsing Hua University, Hsinchu, Taiwan in 2011. He is currently a post doc in National Tsing Hua University. His research interests include data mining, neural networks, and quality engineering.