

Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization

Hany M. Harb

Computers and Systems Engineering Dept.
Faculty of Eng., Azhar University
Cairo, Egypt

Abeer S. Desuky

Mathematics Dept.
Faculty of Science, Azhar University
Cairo, Egypt

ABSTRACT

Classification analysis is widely adopted for healthcare applications to support medical diagnostic decisions, improving quality of patient care, etc. A subset dataset of the extensive amounts of data stored in medical databases is selected for training. If the training dataset contains irrelevant features, classification analysis may produce less accurate and less understandable results. Feature subset selection is one of data preprocessing step, which is of immense importance in the field of data mining. This paper proposes the filter and wrapper approaches with Particle Swarm Optimization (PSO) as a feature selection methods for medical data. The performance of the proposed methods is compared with another feature selection algorithm based on Genetic approach. The two algorithms are applied to three medical data sets. The results show that the feature subset recognized by the proposed PSO when given as input to five classifiers, namely decision tree, Naïve Bayes, Bayesian, Radial basis function and k-nearest neighbor classifiers showed enhanced classification accuracy over all given types of classification methods.

Keywords

Feature selection, Particle Swarm Optimization, medical datasets, Decision tree, Naïve Bayes, Bayesian Classifier, Radial Basis Function, K-Nearest Neighbor.

1. INTRODUCTION

The cost of medical and healthcare is increasing more rapidly than the readiness and the ability to pay for it. At the same time, due to the availability of computers, more and more data is becoming accumulated. Such a large amount of data cannot be processed by the experts in a short time to make diagnosis, prognosis and treatment schedules in short time. As a result, data mining has become critical to the medical healthcare world [1] [2].

A significant step in the data mining is data preprocessing, since the quality of decisions is based on the quality of data. Enhancing the medical database improves the quality of medical diagnosis. Data preprocessing steps are data cleaning, data integration, data transformation and data reduction (feature subset selection). Some attributes of datasets may be redundant as their information may be contained in other attributes. Some extra attributes can increase computation time having an impact on the diagnosis accuracy. Some data in the dataset may not be useful for diagnosis and thus can be eliminated before learning.

The goal of feature selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes [1]. Data reduction reduces the number of features, and removes irrelevant, redundant, or noisy data. This reduction has great immediate effects on speeding up data mining algorithm, and improving mining performance such as predictive accuracy and result comprehensibility [3].

In this work we present use of multivariate filters and wrapper approach, which uses PSO with CFS and PSO with the classifier used in the classification process. The resulted subset of features is provided as input to five classifiers. Our work is compared with the work of another researcher in [2] who used GA with CFS as fitness evaluator and the results show more enhanced classification by providing the features selected by our proposed approaches.

2. THE FEATURE SELECTION

Extracting useful data from arbitrarily large data collections or data streams is now of special interest within the data mining community. Researchers and practitioners realize that the feature selection is an integral component to successful data mining.

Feature selection, as a process of selecting a subset of original features according to certain criteria, is an important and frequently used as a reduction technique for data mining. Feature selection has been an active research area for decades in fields such as machine learning and data mining.

A typical feature selection process consists of four basic steps (shown in Figure 1), namely, subset generation, subset evaluation, stopping criterion, and result validation. Subset generation is a search procedure that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous one. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied, then the selected best subset usually needs to be validated by prior knowledge or different tests via synthetic and/or real world data sets.

Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression. Feature selection algorithms designed with different evaluation criteria broadly fall into three categories [5]: the filter, wrapper, and hybrid models. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm.

The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming for improving mining performance, but it also tends to be more computationally expensive than the filter model.

The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages. Most machine learning algorithms choose the most appropriate attributes to be promising attribute to split on at each point, and theoretically, should never select irrelevant or unhelpful attributes. Having more features should surely—in theory—result in more discriminating power, but in practice, adding irrelevant or distracting attributes to a dataset often “confuses” machine learning system where learning models tend to over fit and become less comprehensible.

Machine learning algorithms can be used for attribute selection. For instance, a decision tree algorithm is firstly applied to the full dataset, and then only those actually used attributes in the tree are selected. Although this selection would have no effect at all if the second stage merely built another tree, it will have an effect on a different learning algorithm.

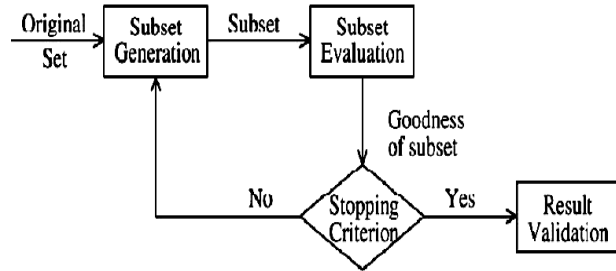


Fig 1: The feature selection process

For example, the nearest-neighbor algorithm is notoriously susceptible to irrelevant attributes, and its performance can be improved by using a decision tree builder as a filter for attribute selection first. The resulting nearest-neighbor method can also perform better than the decision tree algorithm used for filtering [5] [6].

3. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) is an evolutionary computation technique proposed by Kennedy and Eberhart in 1995 [7]. In PSO, a population, called a swarm, of candidate solutions are encoded as particles in the search space. PSO starts with the random initialization of a population of particles. Each particle has its objective function value which is decided by a fitness function. The particles fly in the search space with a velocity adjusted by each particle own flying memory and companion's flying experience [8].

During movement, the current position of particle i is represented by a vector $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, where D is the dimensionality of the search space. The velocity of particle i is represented as $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})$ and it must be in a range defined by parameters v_{min} and v_{max} . The personal best position of the particle $localbest$ is the best previous position of that particle and the best position obtained by the population thus far is called $globalbest$.

According to the following equations, PSO searches for the optimal solution by updating the velocity and the position of each particle:

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

$$v_{id}(t+1) = w * v_{id}(t) + c_1 * r_{1i} * (p_{id} - x_{id}(t)) + c_2 * r_{2i} * (p_{gd} - x_{id}(t)) \quad (2)$$

where t denotes the t^{th} iteration, $d \in D$ denotes the d^{th} dimension in the search space, w is inertia weight, c_1 and c_2 are the learning factors called, respectively, cognitive parameter, social parameter, r_{1i} and r_{2i} are random values uniformly distributed in $[0, 1]$, and finally p_{id} and p_{gd} represent the elements of $localbest$ and $globalbest$ in the d^{th} dimension. The personal best position of particle i is calculated as

$$y_i(t+1) = \begin{cases} y_i(t) & \text{if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & \text{if } f(x_i(t+1)) < f(y_i(t)) \end{cases} \quad (3)$$

4. THE PROPOSED METHOD

4.1 Filter Approach (PSO-CFS)

Generally, the filter approaches are independent of the learning induction algorithm. Filters estimate a relevance index for each feature to measure how relevant a feature is to the target, then rank features by their relevance indices and perform search according to the ranks or based on some statistical criterion [2].

In this work, PSO is proposed as a filter with CFS (Correlation-based Feature Selection) as a fitness function. Like the majority of feature selection techniques, CFS uses a search algorithm along with a function to evaluate the worth of feature subsets. CFS measures the usefulness of each feature for predicting the class label along with the level of inter correlation among them, based on the hypothesis: Good feature subsets contain features highly correlated (predictive of) with the class, yet uncorrelated with (not predictive of) each other [9]. CFS can be combined with search strategies such as forward selection, backward elimination, bi-directional search, genetic search and best-first search. In this paper we have PSO as search method with CFS as fitness function. It is then compared to the work in [2] where GA used as search method with CFS as fitness function.

4.2 Wrapper Approach

Another FS approach is wrapper approach. Instead of ranking every individual feature, wrappers rank feature subsets by the prediction performance of a classifier on the offered subset. Unlike filters, wrappers can be used to search through all possible subsets of features and explore the mutual information between features. After choosing a classifier, wrappers evaluate the classification performance either by cross-validation or theoretical performance bounds. We also used wrapper with PSO for improving the prediction performance of each used classifier.

4.3 The Medical Datasets

The Breast Cancer Dataset was obtained from the University Medical Centre, Institute of Oncology. It includes 286 instances, 201 instances of one class and 85 instances of another class. Instances are described by 9 attributes, some of which are nominal and some are linear and the output class to be predicted is recurrence-events or no-recurrence-events.

The Heart Statlog dataset consist of 270 instances are described by 13 input attributes and the output class to be predicted is presence or absence of heart disease.

The multi class Dermatology dataset consists of 366 instances, 34 inputs and the output class to be predicted has 6 class labels. The differential diagnosis of erythematous-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are seborrheic dermatitis, psoriasis, pityriasis rosea, lichen planus, pityriasis rubra pilaris, and chronic dermatitis. This database contains 34 attributes, one of them is nominal and the rest are linear valued [9].

5. RESULTS

The multivariate filter: PSO with Correlation based feature selection as subset evaluating mechanism has been used with three medical datasets from the UCI machine Learning Repository: Breast Cancer, Heart Statlog and Dermatology dataset as a part of feature selection step.

For PSO, population size is 20, number of iterations is 20, individual weight is 0.34 and inertia weight is 0.33. Tables 1:3 show the number of features selected by proposed filter PSO with CFS for the three medical dataset and they also show the number of features selected by proposed filter PSO applied on the five Weka classifiers Naïve bayes, Decision tree C3.4, RBF, K-NN and Bayesian classifier. The five classifiers have been tested on the medical datasets using the relevant feature as identified by the proposed filters. K-NN was experimented with different values of K neighbors that have been used in [2].

Table 1. Classification performance for Breast Cancer dataset

Classifier	Approach	# features	Accuracy	rmse
Naïve Bayes	PSO+NBay.	4	75.52	0.44
	PSO+CFS	5	74.13	0.45
	GA+CFS	5	72.16	---
	All Inputs	9	71.13	0.48
Bayesian	PSO+Bay.	3	73.08	0.43
	PSO+CFS	5	73.43	0.45
	GA+CFS	5	70.10	0.48
	All Inputs	9	70.10	0.49
RBF	PSO+RBF	4	76.22	0.43
	PSO+CFS	5	72.03	0.44
	GA+CFS	5	70.10	0.45
	All Inputs	9	68.04	0.47
Decision Tree	PSO+DT	5	74.13	0.44
	PSO+CFS	5	72.03	0.44
	GA+CFS	5	66.00	0.48
	All Inputs	9	68.04	0.49
K-NN	PSO+KNN	5	76.22	0.43
	PSO+CFS	5	74.83	0.43
	GA+CFS	5	74.50	0.45
	All Inputs	9	70.10	0.45

Table 1 illustrates the improvement in classification accuracy and root-mean-square error (RMSE) of the five classifiers on the Breast Cancer dataset as result of feature selection. The number of features selected by our proposed wrapper model for three classifiers - Naïve bayes and RBF classifier- is fewer

number of features while keeping higher classification accuracy. Classification accuracy of Bayesian classifier is the same for our two proposed models which is higher than GA_CFS. Despite the elimination of 6 features out of 9 features.

For the heart statlog dataset (Table 2), the classification accuracy of KNN classifier remained the same with all inputs as well as with features selected by proposed filter PSO - CFS, illustrating the fact that elimination of 6 irrelevant features has not get the classification accuracy worse. Exceptional case exists here where the classification accuracy obtained with GA_CFS for KNN classifier has exceeded our proposed wrapper model by 2%.

Table 2. Classification performance for Heart dataset

Classifier	Approach	# features	Accuracy	rmse
Naïve Bayes	PSO+NBay.	7	85.56	0.35
	PSO+CFS	7	85.19	0.36
	GA+CFS	7	84.78	0.36
	All Inputs	13	83.70	0.37
Bayesian	PSO+Bay.	5	84.82	0.35
	PSO+CFS	7	84.44	0.36
	GA+CFS	7	82.16	0.35
	All Inputs	13	82.61	0.36
RBF	PSO+RBF	9	85.19	0.35
	PSO+CFS	7	83.70	0.35
	GA+CFS	7	83.70	0.36
	All Inputs	13	82.61	0.37
Decision Tree	PSO+DT	4	83.33	0.37
	PSO+CFS	7	80.74	0.39
	GA+CFS	7	76.08	0.45
	All Inputs	13	76.09	0.46
K-NN	PSO+KNN	7	83.70	0.37
	PSO+CFS	7	82.59	0.37
	GA+CFS	7	85.87	0.37
	All Inputs	13	82.60	0.37

Table 3 shows the classification accuracy and the RMSE for the Dermatology dataset. The results show that using feature subset selection enhances the classification accuracy and the RMSE of all the five classifier for Dermatology dataset. Only the RBF classifier performance has not enhanced accuracy with our proposed models against the enhancement with GA_CFS.

The main objective of this paper is to present the significance of the feature selection. The experiment results obviously show a perceivable improvement in classification accuracy for all used classifiers except for Bayesian classifier and show that the reduction in irrelevant attribute has not decreased the performance of a classifier.

Table 3. Classification performance for Dermatology dataset

Classifier	Approach	# features	Accuracy	rmse
Naïve Bayes	PSO+NBay.	22	99.18	0.47
	PSO+CFS	20	99.45	0.05
	GA+CFS	21	98.39	0.05
	All Inputs	34	97.58	0.07
Bayesian	PSO+Bay.	23	99.45	0.04
	PSO+CFS	20	99.45	0.04
	GA+CFS	21	99.13	0.05
	All Inputs	34	99.13	0.05
RBF	PSO+RBF	15	89.62	0.2
	PSO+CFS	20	86.07	0.22
	GA+CFS	21	98.39	0.07
	All Inputs	34	82.79	0.11
Decision Tree	PSO+DT	14	97.27	0.09
	PSO+CFS	20	97.54	0.08
	GA+CFS	21	97.58	0.1
	All Inputs	34	94.35	0.14
K-NN	PSO+KNN	22	98.63	0.09
	PSO+CFS	20	97.27	0.11
	GA+CFS	21	97.58	0.09
	All Inputs	34	95.96	0.10

6. CONCLUSIONS

The proposed models have been experimented with three medical datasets. The experimental results clearly illustrate that the filter PSO_CFS improves classification accuracy of the applied classifiers for all the medical dataset used. The proposed wrapper technique –PSO applied to each classifier– show the best classification accuracy.

For the dermatology dataset, the relevant attributes identified by proposed models have improved classification accuracy of Naïve bayes, Decision tree, RBF and K-NN. Further the performance of Bayesian slightly increased with the relevant attributes as input by proposed models. Just two exceptional cases have been found when GA_CFS has better accuracy

than our methods: the first case with KNN classifier when applied on heart statlog dataset, and the second with RBF

7. REFERENCES

- [1] Adam Woznica, Phong Nguyen, Alexandros Kalousis, "Model mining for robust feature selection", KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, PP 913-921, 2012.
- [2] Asha Gowda Karegowda, M.A.Jayaram, A.S .Manjunath, "Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning", International Journal of Computer Applications (0975 – 8887), Vol. 23– No.2, June 2011.
- [3] MIT Lincoln Laboratory: <http://www.ll.mit.edu/IST/idaval/>.
- [4] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings Volume 10: 4-13, The Fourth International Workshop on Feature Selection in Data Mining, Hyderabad, India, June 21st, 2010.
- [5] Ian H.Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers, Elsevier Inc. 2005.
- [6] Hany M. Harb, Afaf A. Zaghot, Mohamed A. Gomaa and Abeer S. Desuky, "Selecting Optimal Subset of Features for Intrusion Detection Systems", Advances in Computational Sciences and echnology, Research India Publications, Volume 4 Number 2, pp. 179-192, 2011.
- [7] XUE, Bing; ZHANG, Mengjie; BROWNE, Will N., "Multi-objective particle swarm optimisation (PSO) for feature selection", In: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference. ACM, pp. 81-88, 2012.
- [8] LIU, Yuanning, et al. An improved particle swarm optimization for feature selection. Journal of Bionic Engineering, 8.2: 191-200, 2011.
- [9] Hall, Mark A and Smith, Lloyd A, " Feature subset selection: a correlation based filter approach", Springer, 1997.