Biomedical Research 2017; 28 (11): 4749-4751 *ISSN 0970-938X*
www.biomedres.info

Abnormality detection using weighed particle swarm optimization and smooth support vector machine.

Latchoumi TP[1], Latha Parthiban[2*]

[1]Research Scholar, Department of Computer Science and Engineering, Sathyabama University, Assistant Professor, Vignan's University, Vadlamudi, Andra Pradesh, India

[2]Department of Computer Science, Pondicherry University CC, Pondicherry, Tamil Nadu, India

Abstract

In this paper, a new hybrid classification approach, which uses Weighted-Particle Swarm Optimization (WPSO) for data clustering in sequence with Smooth Support Vector Machine (SSVM) for classification is proposed. The performance of WPSO clustering is compared with K means and fuzzy methods using intercluster, intracluster and validity index. The accuracy of proposed WPSO-SSVM classification methodology are 83.76% for liver disorder, 98.42% for WBCD, 95.21% for mammographic mass data which are better than in existing literature.

Keywords: Smooth support vector machine (SSVM), Particle swarm optimization (PSO), Clustering, Classification.

Accepted on March 14, 2017

Introduction

Medical data mining has a great potential for exploring hidden patterns and extracting useful information for decision support [1]. Benefits of introducing machine learning into medical analysis are to increase diagnostic accuracy, reduce costs and human resources [2]. Case based reasoning [3] process is an approach for developing knowledge-based medical decision support system which solves new problems based on the solutions of similar past problems.

Materials and Methods

Assume a medical library with each case in the library as index of corresponding features (e_1, e_2, ..., e_N) having an associated action, with collection of features F_j ($j=1.... n$) representing the cases and variable V denoting the action. The i^{th} case e_j in the library can be represented as an $n+1$-dimensional vector, i.e. $e_i=(x_{i1}, x_{i2}, ..., x_{in}, y_i)$. Where x_{ij} corresponds to the value of feature F_j ($j=1... n$) and y_i corresponds to the action ($i=1... n$). If for each j ($1 \leq j \leq n$) a weight w_j (w_j (0, 1)) has been assigned to the j^{th} feature to indicate the importance of the feature, then for any pair of cases e_p and e_q in the library, a weighted distance metric $d_{pq}^{(w)}$ is defined as:

$$d_{pq}^{(w)} = d^{(w)}\left(e_p, e_q\right) = \sum_{j=1}^{n} w_j^2 \quad (x_{pj} - x_{qj})^{2^{1/2}}$$

Where x_{pj} is the p^{th} case with j^{th} feature and x_{qj} is the q^{th} case with j^{th} feature. Using the weighted distance a similarity measure $SM_{pq}^{(w)}$ is calculated using $SM_{pq}^{(w)}=1/(1+\alpha d_{pq}^{(w)})$

Where α is a positive parameter. The weighted feature assignment algorithm is presented in Figure 1.

Input: A set of data samples.

Output: A set of weighed attribute samples.

1. Initialize w_j with random values in [0, 1].
2. Compute w_j for each j using $\Delta w_j = -\eta \frac{\partial E}{\partial w_j}$ (Assume α =0.6 and η=0.05)
3. Calculate evaluation function $\mathbf{E(w)} = \mathbf{2} \times \left[\frac{\sum_{pq}\sum_{q<p} SM_{pq}^{(w)}\left(1-SM_{pq}^{(1)}\right)+SM_{pq}^{(1)}\left(1-SM_{pq}^{(w)}\right)}{N(N-1)}\right]$ where SM_{pq} is the similarity measure and N is the number of data in dataset.
4. Update w_j with $w_j + \Delta w_j$ for each j.
5. Compute w_j and E (w) until convergence (E becomes less than or equal to a given threshold or until the number of iterations exceeds a certain predefined number).
6. Use final w_j as the weighted feature for PSO-clustering.

Figure 1. *Weighed feature assignment algorithm.*

PSO is a population-based search algorithm and each particle is associated with a velocity and its algorithm is presented in Figure 2.

A nonlinear version of the SSVM [4] is used for classification of datasets after clustering.

Results

The WBCD, mammographic mass and liver disorder dataset are obtained from UCI machine learning repository [5]. Weighed PSO clustering is applied on the datasets (Figure 3) and compared with K-means and FCM in terms of intercluster, intra cluster and validity index as shown in Table 1. The inter cluster distance of any two cluster should be high which is best for PSO as seen in Table 1. Intra cluster means the compactness of a cluster and its value should be least as possible and is again best for PSO.

The clustered output is classified using SSVM using fivefold cross validation in which randomly split database is averaged to provide the best indication of true classification performance and the performance comparison of datasets is presented in Table 2 and accuracy is shown in Figure 4.

Input: A set of weighed attribute samples.

Output: PSO clustered results.

1. Initialize a particle with random cluster centroid where each particle is collection of cluster centers.
2. for t=1 to t_{max}
   for each particle
   for each data vector Z_p in medical data set
   calculate d (m_j, Z_p) for all cluster centroids using f

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} w_j (z_{pk} - m_{jk})^2}$$

   where m_j denotes centroid vector of cluster j
   Z_p denotes the p^{th} data vector
   N_d denotes the number of input dimension
   w_j denotes the weight of each feature
   Assign Z_p to a cluster C_i which has minimum Euclidean distance
3. Calculate the fitness of particle using the fitness function F

$$F = \left[\frac{\sum_{j=1}^{N_c} \left[\sum_{z_p \in C_{ij}} \sqrt{\frac{\sum_{k=1}^{N_d} w_j (z_{pk} - m_{jk})^2}{C_{ij}}} \right]}{N_c} \right]$$

   Z_p denotes the p^{th} data vector
   $|C_{ij}|$ is the number of data vectors belonging to cluster C_{ij}
   N_c denotes the number of cluster centroids
   N_d denotes the input dimension
4. Update the global best and local best as follows:
   do
   for each particle
   calculate fitness function F
   if fitness value is better (minimum) than best fitness pbest
   set current value as the new pbest
   end
   Gbest=min (Pbest).
5. Update the particle velocity and location.

Figure 2. *PSO based clustering algorithm.*

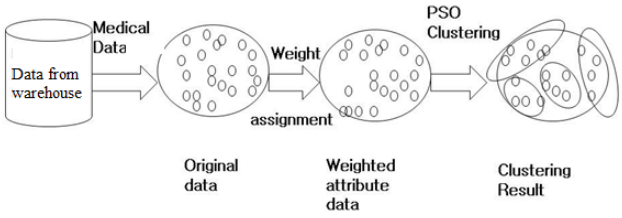

Figure 3. *Weighed PSO based clustering.*

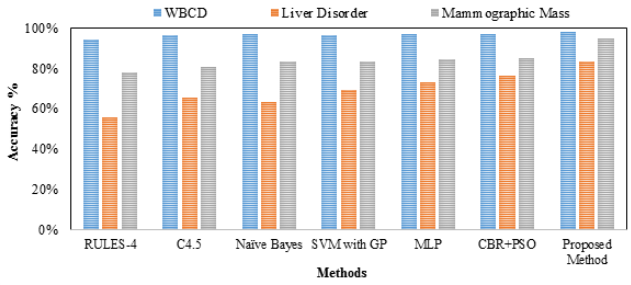

Figure 4. *Accuracy (%) comparison of proposed method with other existing methods in literature.*

Table 1. *Comparison of inter, intra and validity index with FCM, K-means and PSO for breast cancer (WBCD) and Liver disorder dataset [6].*

Measures	FCM			K-means			PSO		
	WBCD	Liver disorder	Mammographic mass	WBCD	Liver disorder	Mammographic mass	WBCD	Liver disorder	Mammographic mass
Inter Cluster	708.56	87.4948	24.43	713.944	109.817	23.91	941.771	172.3005	237.639
Intra Cluster	NA	NA	NA	11.4572	2.6474	0.3942	0.292068	0.0846	0.004
Validity Index	NA	NA	NA	0.016047	0.0241	0.0164	0.00031	0.0049	0.0001

Table 2. *Performance comparison of datasets.*

Methods	WBCD			Liver Disorder			Mammographic Mass		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
RULES-4	94.74%	96.43%	92.56%	55.90%	56.78%	54.57%	78.13%	79.55%	75.67%
C4.5	96.80%	97.12%	94.54%	65.59%	66.78%	64.85%	81.13%	84.54%	79.56%
Naive Bayes	97.36%	98.53%	96.23%	63.39%	66.45%	61.23%	83.43%	86.64%	82.36%
SVM with GP	96.70%	98.40%	94.97%	69.70%	71.67%	65.67%	83.66%	85.54%	81.14%
MLP	97.20%	98.57%	96.25%	73.05%	74.57%	72.46%	84.79%	87.64%	82.45%
CBR+PSO	97.41%	98.53%	96.45%	76.81%	77.67%	73.68%	85.29%	87.64%	83.44%
Proposed method	98.42%	99.38%	97.35%	83.16%	86.16%	77.17%	95.21%	97.57%	93.45%

Conclusions

This paper proposes a new WPSO-SSVM technique to improve the classification accuracy of medical datasets and the obtained results are found to outperform all the present state-of-art classifiers existing in literature. The future work will be to test the proposed technique in other benchmark datasets to prove the robustness of the proposed algorithm.

References

1. Pei CC, Jyun JL, Chen HL. An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. Comp Met Prog Biomed 2011; 1-11.
2. Bojarczuk HSL, Freitas AA. Genetic programming for knowledge discovery in chest-pain diagnosis: exploring a promising data mining approach. IEEE Eng Med Biol Magaz 2000; 19: 38- 44.
3. Klaus DA. Case-based reasoning for medical decision support tasks: the inreca approach. Artificial Intel Med 1998; 12: 25-41.
4. Yuh-Jye L, Mangasarian. Smooth support vector machine. Comp Optimiz Appl 2001; 20: 5-22.
5. Data Mining Institute, University of Wisconsin, Technical Report 99-03.
6. https://archive.ics.uci.edu/ml/datasets.html (Liver disorders, Mammographic mass and Wisconsin Breast Cancer (Original)).

*Correspondence to

Latha Parthiban

Department of Computer Science

Pondicherry University CC

Pondicherry

Tamil Nadu

India