Medical Image Retrieval System Using PSO for Feature Selection

Sasi Kumar. M\(^1\) and Dr. Y.S. Kumaraswamy\(^2\)

Abstract—Content-based image retrieval (CBIR) is a widely researched area, with various techniques proposed in literature for feature extraction, classification and retrieval. But, when database size increases, overall retrieval performance deteriorates significantly. Features in pattern recognition are individual measurable heuristic properties of the image under observation. Choosing discriminating/independent features is the key for the efficiency of pattern recognition algorithms to succeed in classification. In this paper, Information Gain is used to achieve the structure of a feature sets to find a subset of the original feature vector for efficient computation. The obtained features are optimized using Particle Swarm Optimization (PSO).

Keywords—Content-based image retrieval, Information Gain, Particle swarm optimization, Multilayer Perceptron.

I. INTRODUCTION

Content-based image retrieval (CBIR) is a widely researched area, with various techniques proposed in literature for feature extraction, classification and retrieval. But, when database size increases, overall retrieval performance deteriorates significantly. For content-based image classification and retrieval, main issues are

1) How to select features to achieve highest discrimination,
2) Combining them effectively,
3) Application of proper distance metrics,
4) Location of optimal classifier configuration for classification problems,
5) scaling/adapting classifier when many classes/features are incrementally introduced and finally,
6) Training classifier to maximize classification accuracy.

Current state-of-the-art classifiers like SVMs, Bayesian, Artificial Neural Networks (ANNs), etc. are unable to handle such requirements as a single classifier, even if they are powerful and well-trained. They are unable to discriminate classes efficiently over indefinitely large feature sets.

In many problems, dimension reduction is essential before data analysis is performed. The usual parameter for dimension reduction is a desire to preserve relevant information of original data based on optimality criteria.

In pattern recognition and general classification problems, methods including Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Fisher Linear Discriminate Analysis (LDA) were extensively used where such methods found mapping from the original feature space to a lower dimensional feature space.

Some applications might desire to pick a subset of original features rather than finding a mapping which uses all original features. Benefits of locating such features subsets could save the cost of computing unnecessary features, and excluding noisy features and keeping their information using “clean” features. Features in pattern recognition are individual measurable heuristic properties of the image under observation. Choosing discriminating/independent features is key for the efficiency of pattern recognition algorithms to succeed in classification. A bad feature set can deteriorate a good classifier’s performance.

With noise and dimensionality increasing, feature selection and fusion become important steps. A feature that has too much confusing (contradictory) information should be avoided as they confuse classifiers. To reduce data noise, class information correlated weak features must be removed. When class information is missing, a clustering algorithm is used, where cluster scatter/compactness formed with/without a particular feature are compared to decide if a particular feature is redundant. "Curse of dimensionality" is motivation for feature selection. Numerous features increase computational time without major performance change in the testing phase. Over decades, a lot of work was done on feature selection though most algorithms assume selection criterion to be monotonic. This is a major drawback, as really gradual addition/deletion of a feature from a set does not usually change the criterion function value. Fusion of different features generates a new feature set from selected features set. It is unclear how to combine features in the best manner, and how to use the unselected features to ensure that they provide better classification results.

Feature selection is the essence of feature fusion. Feature selection is a well explored problem especially in data mining and CBIR. This issue deals with finding appropriate (suitable or discriminatory) features subset from a given set, for some particular application domain, to the improve accuracy. This includes locating a minimal subset representing the entire set, or to rank the features based on the significance, from the overall set. Filter method [1], wrapper method [2], and hybrid method [2] are the three types of feature selection algorithms. In the first approach, a feature set is evaluated at once.
independent of any clustering algorithm/classifier. The wrapper method needs a clustering algorithm/classifier for each subset evaluation to locate a final subset. Though the filter method is unbiased and fast, better results are obtained from the wrapper method for clustering algorithm or classifier. The Hybrid method is a combination of the filter and wrapper methods.

Many feature selection methods were proposed in the past. The selection method depends on the learning type (supervised or unsupervised) and algorithm used. Most feature selection (evaluation function) methods are used in all algorithms (filter or wrapper), but there are some which are meant exclusively for wrapper methods or filter methods. Four steps describe general feature selection framework, including subset generation/evaluation, stopping criteria, and result validation. In supervised learning, one uses a validation set to check performance/validity of the final set. Subset generation is basically a search algorithm to select a features subset from the original features set. Liu et al. [2] broadly classified all subset evaluation criteria into four types, namely, distance measure, information gain measure, dependency measure, and consistency measure.

This work suggests a computationally efficient system exploiting the structure of a feature sets to find a subset of the original feature vector using information gain. And the obtained features are optimized using Particle Swarm Optimization (PSO). The paper is organized in the following manner: Section 2 reviews some to works available in the literature; section 3 explains the materials and methods used in this research, section 4 gives the experimental results and section 5 concludes the paper.

II. RELATED WORKS

In content-based image retrieval systems, to enhance the retrieval accuracy many methods are framed for the complicated low-level feature extraction and the ‘semantic gap’ reduction among the visual features and the richness of human semantics. Lianze Ma et al., [3] proposed a hybrid model incorporating both Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) for image retrieval and clustering. Using linear/quadratic estimators, the relevance feedback schemes are implemented in content-based image retrieval to considerably advance the retrieval performance. To enhance retrieval performance, an improved PSO method is also proposed with adaptive weight of SVM. To represent a composition of color, luminance, and edge features in the image, to extract low-level feature of an image additionally, a composite histogram scheme is employed. Therefore a composite histogram representation is obtained, as a result. The PSO-SVM approach is experimented and the results reveal that for the problem of image retrieval and clustering, greater recognition accuracy and more recognition speed was achieved.

Umamaheswari et al., [4] explained the process of recognition and classification of brain images like normal and abnormal based on PSO-SVM. For medical diagnosis process, the image classification is emerging as the most important tool. For doctors, to diagnosis the patient regarding to the severity of the diseases, the main requirement in the medical field is the diagnosis of the abnormality of the patient to be classified. The optimal recognition and early detection of diseases are very complicated in case of DICOM images. The aim is to recognize and classify DICOM image based on collective scheme of digital image processing. Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Support Vector Machine (SVM) are employed for optimal recognition and classification. Hence, the collective scheme employing PSO-SVM provides the best approximation capability and faster convergence. Therefore, the proposed PSOSVM is an efficient and effective model.

For two-phase image segmentation, Benoit Mory et al., [5] introduced a new framework called the Fuzzy Region Competition. The idea of Region Competition is that it is the functional present in many preexisting models are introduced to a fuzzy membership function as an enhancement. The convex nature of the new problem and its global solution sets to be stable in thresholding, and also obtains solutions to the equivalent classical formulations in the operation. In the piecewise-constant case, the benefits are revealed. At last medical applications like angiography are motivated and a fast and consistent algorithm is derived to segment images into two non-overlapping smooth regions. It is built around extensive windowing principles and has a specific benefit to feature closed form solutions in each region for the approximation functions based on normalized convolutions, when compared to the preexisting models. In synthetic 2D images and real 3D volumes, the results are revealed.

In traditional CBIR researches, the image relevance evaluation depends on the given criterion. The significant thing is that, in relation to the image database, the ability to change the criterion. Using Particle Swarm Optimization (PSO), Keisuke Kameyama et al., [6] proposed a model to tune the parameters embedded in the relevance evaluation algorithm of a CBIR system, according to the suitability of the retrieved results achieved by optimizing them. To enhance the retrieval ranking score, the parameters that affect the similarity evaluation in a binary shape matching CBIR system was tuned in the experiment. Using PSO, after the tuning, improved according to the given criterion it is observed that the ranking of the retrieved images were enhanced.

III. MATERIALS AND METHODS

A. Information Gain

The main aim of information gain criteria is to discover the amount of unique information is added by a feature to the whole feature set. A features information gain \( f \) can be computed as \( F(S \cup f) - F(S) \), where \( F(.) \) is the evaluation criterion and \( S \) the selected subset of features. The feature with greater information gain is preferred. Bayes error rate, conditional probability, and information gain are a little information gain criteria. Information gain criteria are used in [7-9].
Quinlan [10] suggested a classification algorithm called ID3 that introduced the information gain concept. Information gain is a measure based method, used for selecting best split attributes in decision tree classifiers and indicates the extent to which data’s entropy is reduced. It also identifies values of each particular attribute. Each feature basis gets an information gain value, which is used to decide whether a feature is selected/deleted. Hence, a threshold value for feature selection must be established first; a feature is chosen when its information gain value is bigger than the threshold value.

Let a set of s instances be set A and let B be the set of k classes. Let \( P(B_i, A) \) be the fraction of the examples in A that have class \( B_i \), then, the expected information for the class membership is given by:

\[
Info(A) = -\sum_{i=1}^{k} P(B_i, A) \times \log(P(B_i, A))
\]  

(1)

If a particular attribute X has y distinct values, anticipated information for the decision tree with X as root is the weighted sum of expected information of subsets of X according to distinctive values. Let \( A_i \) be the set of instances whose attribute value of X is \( X_i \).

\[
Info_X(A) = -\sum_{i=1}^{y} \frac{|A_i|}{A} \times Info(A_i)
\]  

(2)

Then, difference between \( Info(A) \) and \( Info_X(A) \) provides information gained by partitioning A according to testing X.

\[
Gain(X) = Info(A) - Info_X(A)
\]

The higher the information gain, the higher the chances of getting pure classes in a target class if the split is based on the variable with the highest gain.

The wrapper model approach depends on feature addition/deletion to compose subset features. The evaluation function is used with a learning algorithm to for subset features estimation. Such an approach is the same as an optimal algorithm searching for optimal results in a dimension space. A subset search is conducted with an optimal algorithm in the wrapper approach, with the subset being evaluated later by a classification algorithm.

**B. Particle swarm optimization (PSO)**

Particle swarm optimization (PSO), a population-based stochastic optimization technique, was developed by Kennedy and Eberhart in 1995 [11]. PSO simulates organisms’ social behavior and is an automatically evolving system. In PSO, every single candidate solution is an "an individual bird in the flock", that is, a particle in the search space. Each particle uses its own memory and the optimal solution is found on the knowledge gained by the swarm. Kennedy and Eberhart introduced a binary version of PSO (BPSO) [11] to solve discrete problems in 1997. In BPSO, each particle shows its position through both binary value \{0\} or \{1\} and velocity is looked upon as a probability change of the particle position. But BPSO has disadvantages similar to evolutionary algorithms. After several generations, these algorithms are easy to trap in a local optimum, which prevents them converging to a global optimal solution. To circumvent premature convergence at a local optimum, Boolean operation was incorporated to create a new \( g_{Best} \) position. The new \( g_{Best} \) replaced the original \( g_{Best} \), ensuring that all particles left the local optimal.

In PSO algorithms, each particle moves within regions of decision space retaining memory of best positions encountered. This best position attained by every swarm particle is informed to all particles. The state-of-the-art hybrid PSO algorithm [12-13] is considered. Particularly, the traditional PSO assumes an n-dimensional search space, \( \subseteq R^n \), where is the total number of decision variables in the optimization problem, and a swarm consisting of N-particles.

Variables are defined as follows in PSO [14]. The position of the particle at time- \( t \) is an \( n \)-dimensional vector denoted by

\[
s_i(t) = (s_{i,1}, s_{i,2}, \ldots, s_{i,n}) \in S.
\]

The velocity of this particle at time- is also an \( n \)-dimensional vector

\[
V_i(t) = (v_{i,1}, v_{i,2}, \ldots, v_{i,n}) \in S.
\]

The best previous position of the particle is a point in S, denoted by

\[
P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,n}) \in S.
\]

The global best position ever attained by all particles is a point denoted by

\[
P_{gb} = (p_{gb,1}, p_{gb,2}, \ldots, p_{gb,n}) \in S.
\]

Then, the PSO assumes that the swarm is manipulated by equations

\[
V_{i}(t+1) = k \cdot \left[ w(t) \cdot V_{i}(t) + c_1 \cdot rand_1 \cdot (P_{i} - S_i(t)) + c_2 \cdot rand_2 \cdot (P_{gb} - S_i(t)) \right]
\]

and

\[
S_i(t+1) = S_i(t) + V_i(t+1)
\]

(3)

and

(4)

where \( i= 1, \ldots, N \); \( c_1 \) and \( c_2 \) are cognitive and social parameters, respectively; \( r_1 \) and \( r_2 \) are uniformly distributed random numbers within \( (0, 1) \).

The inertia weighting factor for particle velocity - is defined by the inertial weight approach

\[
w(t) = \frac{\text{W}_{\text{max}} - \text{W}_{\text{min}}}{t_{\text{max}}} \cdot t
\]

(5)

where are the maximum iterations, and is the present number of iterations; and are the upper and lower limits of inertia weighting factor, respectively. Moreover, to guarantee convergence of PSO algorithm, the constriction factor is defined as [15],

\[
k = \frac{2}{2 - \varphi - \sqrt{\varphi^2 - 4\varphi}}, \varphi = c_1 + c_2, \varphi \geq 4.
\]

(6)
C. Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is a preferred supervised learning network model. It consists of an input layer, one or more hidden layer and an output layer. Nodes from a given layer are connected to all neurons in the next layer. During the training phase, scalar weight in each connection is adjusted. During training, examples are fed, and weights for the predicted outputs are computed. Error is obtained by comparing the output with the target output and it is propagated back through the network, and the weights are adjusted.

The proposed neural network Fuzzy Softmax Multi-Layer Perceptron (FS-MLP) [16] Neural Network improves the classification accuracy of traditional MLP Neural Network model by introducing a fuzzy hidden softmax layer. The construction of the proposed model is given in table 1.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>PARAMETERS USED IN THE PROPOSED FS-MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Neuron</td>
<td>50</td>
</tr>
<tr>
<td>Output Neuron</td>
<td>2</td>
</tr>
<tr>
<td>Number of Hidden Layer</td>
<td>2</td>
</tr>
<tr>
<td>Transfer function of first hidden layer</td>
<td>Fuzzy Softmax</td>
</tr>
<tr>
<td>Learning Rule of first hidden layer</td>
<td>Levenberg-Marquardt</td>
</tr>
<tr>
<td>Transfer function of second hidden layer</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>Learning rule of second hidden layer</td>
<td>Levenberg-Marquardt</td>
</tr>
</tbody>
</table>

The activation function contains the output value such that it is in the range of 0 and 1 or -1 to 1. Mathematically the interval activity of the neuron can be shown to be:

$$v_k = \sum_{j=1}^{p} w_{jk} x_j$$

(7)

Where $x_i$ is the input and $w_{jk}$ is the weights. The output of the neuron $v_k$ would therefore be the outcome of some activation function on the value of $v_k$. The most common type of activation used to construct the neural network is the sigmoid function.

A sigmoid activation function uses the sigmoid function to determine its activation. The sigmoid function is given as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

(8)

This function can range between 0 and 1, but it is also sometimes useful to use the -1 to 1 range. An example of the sigmoid function is the hyperbolic tangent function where the range is -1 to 1.

$$\varphi(v) = \tanh\left(\frac{v}{2}\right) = \frac{1 - \exp(-v)}{1 + \exp(-v)}$$

(9)

The softmax activation function [17] make sure that the outputs adapt to the mathematical requirements of multivariate classification probabilities [18]. If the classification problem has C classes, then each class is modeled by one output in the network. If $Z_i$ is the weighted sum of products between its weights and inputs for the i-th output, i.e.,

$$Z_i = \sum_j w_{ji} y_{ji}$$

(10)

Then

$$\text{softmax}_i = \frac{e^{Z_i}}{\sum_{j=1}^{C} e^{Z_j}}$$

(11)

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>CLASSIFICATION ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without IG-PSO feature selection</td>
<td>With IG-PSO feature selection</td>
</tr>
<tr>
<td>MLP</td>
<td>93.1</td>
</tr>
<tr>
<td>Proposed FS-MLP</td>
<td>96.55</td>
</tr>
</tbody>
</table>

Fig. 1 Sample images used in this work

As the softmax activation function ensures that all outputs conform to the requirements for multivariate probabilities. That is,

$$0 < \text{softmax}_i < 1$$

for all $i=1,2,\ldots,C$ and

$$\sum_{i=1}^{C} \text{softmax}_i = 1$$

IV. RESULTS AND DISCUSSION

For evaluating the proposed method, a dataset of 58 medical images with four class labels were used in the experimental setup. Information gain is used to select the top 50 relevant
attributes. Figure 1 shows some of the MRI images used in this work.

The results obtained from regular MLP Neural Network and the proposed FS-MLP Neural Network with and without the proposed IG-PSO feature selection are given in Table 2 and is shown in Figure 2.

![Classification accuracy measured in percentage.](image)

Fig. 2   Classification accuracy measured in percentage.

V. CONCLUSIONS

In this paper, it was proposed to use a combination of information gain and PSO for feature selection. The extracted features using the proposed method were trained with the existing MLP Neural network classifier and compared with the proposed FS-MLP neural network. The classification accuracy of 98.27% was achieved using the proposed method. And also the proposed feature selection decreases the overall processing time for a given query.

REFERENCES


M. Sasi Kumar obtained Master of Computer Applications(MCA) from Bharathidasan University, Tamil Nadu,India and Master of Engineering(M.E)- CSE from Sathyabama university, Tamil Nadu,India. He is also pursuing Ph.D in Computer Science & Engineering, Sathyabama University, Tamil Nadu,INDIA.

Dr .Y.S .Kumaraswamy, working presently as Sr.Prof and Head in Department of MCA (VTU) at Dayananda Sagar of College of Engineering, Bangalore. He has published more than 150 research papers and guiding Ph.D students in the area of Computer Science/ Mathematics and also selection committee member for ISRO/UGC/DSI profiles