VOL. 12, NO. 9, MAY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ADAPTIVE GENETIC ALGORITHM BASED FUZZY SUPPORT VECTOR MACHINE (AGA-FSVM) QUERY MECHANISM FOR IMAGE MINING

B. Meena Preethi[1] and P. Radha[2]
[1]Department of BCA and M.Sc. SS, Sri Krishna Arts and Science College, Coimbatore, India
[2]PG and Research Department of Computer Science, Government Arts College, Coimbatore, India
E-Mail: meenapreethibphd@gmail.com

ABSTRACT

Big data is the buzz word and its implication is now emerging in the medical field. In order to offer good patient care, medical experts tend to test certain hypotheses by querying huge volumes of unstructured medical data. In this research article, structured image data of the epilepsy patients are obtained from the dataset. Feature extraction process is carried out using Adaptive Genetic Algorithm (AGA). Next image querying of Epilepsy patients is carried out using Fuzzy Support Vector Machine (FSVM). Experiments are carried out using MATLAB. Two types of criteria are used to validate the proposed AGA-FSVM mechanism such as accuracy of fulfilling an advanced medical query and the efficiency in terms of retrieval time. Simulation results shows that the proposed AGA-FSVM attains better accuracy with less computation time.

Keywords: big data, image mining, medical data, image querying, genetic algorithm, support vector machine.

INTRODUCTION

Big data is most commonly defined using the ‘‘3Vs model” namely Volume, Velocity and Variety. Volume denotes to the large amount of data. Velocity meant for the speed at which this data is generated, captured and analyzed. Variety shows the heterogeneity of data. Conventional Relational Database Management Systems (RDBMs) is not capable enough to handle data of this massive size with varying data types and generated at this fast rate. Recent advancements in image acquisition and storage technology have led to fabulous expansion in notably large and detailed image databases particularly in the Big Data form. Hence image mining in big data deals with the extraction of implied knowledge, image data relationship, or other patterns that are not explicitly stored in the image databases. It is notable that image mining is unlike from low-level computer vision and image processing techniques for the reason that the focus of image mining is in extraction of patterns from large collection of images, whereas the focus of computer vision and image processing techniques is in understanding and/or extracting specific features from a single image. This paper proposes Adaptive Genetic Algorithm based Fuzzy Support Vector Machine (AGA-FSVM) Query Mechanism for Image Mining in large image datasets.

RELATED WORKS

Feature selection has been one of the key components of many pattern recognition systems such as image classification [1, 2] and cancer classification systems [3], when more and more diverse information is available to characterize entities such as images and objects to be classified. Since more features do not always lead to better classification performance, feature selection aims to identify a set of relevant and necessary features and to reduce the dimensionality of feature space for improving classification performance [4]. It will also reduce storage and computational costs. There are two types of feature selection methods: filters and wrappers[5, 6]. Filter type methods utilize the general characteristics of feature data and select the top-ranked features according to a criterion. In general, they aim to maximize the relevancy of a set of features while minimizing the redundancy among features [3, 7, 8, 9].

Low level visual features such as the color, the texture and the shape are fundamental for characterizing the visual content such as images [10, 11, 12, 13]. In this paper, we investigate 75 descriptors of these three types of visual features which have been widely used and popular for the visual content representation in various tasks such as the image classification and the image retrieval. Color features include the color histogram [14, 14], the dominant color [16], color moments [17], the color set [18], the color structure descriptor [16], the color layout [16], and the scalable color descriptor [19]. In combination with different color spaces and quantization methods, totally 54 color descriptors (indexed from 1 to 54) are extracted for each image.

FEATURE SELECTION USING ADAPTIVE GENETIC ALGORITHM (AGA)

The proposed feature selection method makes use of an adaptive genetic algorithm (AGA) in order to select a subset of visual features from the total features towards achieving the best querying performance from the image dataset. This feature selection mechanism aims to best classify the region of interests (ROIs) represented as $P_x$into their corresponding query labels P_{lb}= $\{y_1, y_2, \ldots, y_L\}$ with a subset of M visual features.

Adaptive Genetic Algorithm

The proposed AGA algorithm contains five portions namely the encoding strategy, the observation operator, the fitness function, the improved rotation gate for the mutation, and the quantum crossover. It is noteworthy that selection is performed on individual visual features, as a replacement for individual components of each visual feature.

Encoding strategy

In this work instead of encoding a feature as a binary value (0 or 1) like a traditional genetic algorithm (GA), the adaptive genetic algorithm (AGA) encodes the selection status of a feature with a Q-bit which gives the probability of a feature being selected. Along these lines, the selection task for M features can be encoded as a Q-bits vector with M components (i.e., a chromosome with M genes),

$$q = (q_1, q_2, \ldots, q_M) = \begin{pmatrix} \alpha_1 & \alpha_2 \ldots \alpha_M \\ \beta_1 & \beta_2 \ldots \beta_M \end{pmatrix}, \quad (1)$$

where $q_m = (\alpha_m, \beta_m)^T$ (m = 1, 2, . . . , M) represents the Q-bit (selection status) corresponding to the m-th feature. Here, both α_m and β_m are variables with continuous values in [0, 1], with $|\alpha_m|^2$ being the probability of the m-th feature not being selected, while $|\beta_m|^2 = 1 - |\alpha_m|^2$ being the probability of being selected. With such a coding strategy, the adaptive genetic algorithm searches for the optimal chromosome in terms of the fitness function from the sample space through mutation and crossover operations. Since all α and β in Eq. (1) are of continuous values, the population for quantum chromosome sample space is much larger than that for the traditional binary valued chromosome sample space. Consequently, it will be of greater probability using AGA to find a globally optimal chromosome in such a space.

Observation operator

For completing the task of feature selection, the chromosome need to be changed usual binary bits for pointing whether the corresponding features are selected or not. In our AGA approach, such a conversion is habitually referred to as the inspection course of action. In this research work, the inspection course of action is formulated as:

$$o_m = \begin{cases} 0, \text{if } |\alpha_m|^2 > r \\ 1, \text{otherwise} \end{cases} \qquad m = 1, 2, \ldots, M, \quad (2)$$

where r is a random number uniformly distributed in [0,1], $o_m = 1$ implies that the m-th feature is selected, otherwise, the m-th feature is discarded.

Fitness function

In our proposed AGA, the fitness function is employed for evaluating the efficacy of a chromosome. As far as feature selection is concerned, an optimal subset of features needs to be chosen / selected that is capable enough to attain the best querying accuracy. That's why, the fitness function fit(o) is defined as the querying accuracy P achieved with the subset of features selected using the chromosome guided selection scheme o,

$$\text{fit}(o) \overset{\Delta}{=} P = \frac{N_{\text{correct}}(o)}{N_{\text{total}}} \quad (3)$$

where $N_{\text{correct}}(o)$ and N_{total} are the number of correctly classified ROIs according to the selection scheme o and the total number of ROIs in the test subset in Dataset, respectively.

Improved rotation gate for mutation

In the proposed AGA, the mutation operation is carried out with respect to the *m*-th gene in a chromosome which is expressed as:

$$\begin{pmatrix} \alpha'_m \\ \beta'_m \end{pmatrix} = \begin{pmatrix} \cos(\Theta_m) & -\sin(\Theta_m) \\ \sin(\Theta_m) & \cos(\Theta_m) \end{pmatrix} \begin{pmatrix} \alpha_m \\ \beta_m \end{pmatrix} \quad (4)$$

where $(\alpha_m, \beta_m)^T$ and $(\alpha'_m, \beta'_m)^T$ are, respectively, the Q-bit of the *m*-th gene before and after the mutation operation, θ_m is the rotation angle and $\theta_m = s(\alpha_m, \beta_m) \cdot \Delta\theta_m$. Here, $\Delta\theta_m$ determines the rotation angle and $s(\alpha_m, \beta_m)$ controls the rotation direction. Upon comprehensive considerations of the relationship between $\Delta\theta$ and evolution generations, and the relationship between $\Delta\theta$ and the fitness values, the rotation angle is alternatively computed as follows:

$$\theta m = s(\alpha_m, \beta m) \cdot \Delta\theta_m \exp - \left\{ \frac{[\text{fit}(o_{\text{opt}}) - \text{fit}(o)]t}{t_{\text{tmax}}} \right\}, \quad (5)$$

where $s(\alpha_m, \beta_m)$ and $\Delta\theta_m$ still represent the rotation angle *t* and $t_{\max}$ are the current and the largest generation numbers, respectively. Hence by this enhanced mutation strategy, the amount of rotation is adaptively adjusted with the evolution generation and the fitness value. Thus, the diversity of the chromosome population is increased, which effectively avoids the premature convergence problem of conventional GA.

Quantum crossover

In this paper, the crossover operation is performed with a probability *Pc* of the population after observation. Suppose that $o^k = (o^k_1, o^k_2, \ldots, o^k_M)$ and $o^k_{cr} = (o^k_{cr,1}, o^k_{cr,2}, \ldots, o^k_{cr,M})$ are the observed states of the *k*-th chromosome before and after crossover operations. The crossover operation can be formulated as:

$$o^k_{cr,m} = o^{((k+m-1))K}_m, \quad (6)$$

where $((k + m - 1))_K = (k + m - 1) \bmod K$, *K* is the number of samples in the population. This operation simulates the interference procedure and can make the best use of the information contained in the population. The crossover operation is very helpful for increasing the diversity of the population and avoiding prematurity of the algorithm.

Algorithm

Input: Feature vector pool $P_{\mathbf{x}}$, half for training and half for testing, query label pool P_{lb},
Crossover probability *Pc* , maximum generation number *tmax*

Output: Feature subset $S_f^{(t)}$ and its indices $I_f^{(1)}$ for the selected features
Initialize: The generation number $t = 0$ and the population $Q = \{\mathbf{q}^1, \mathbf{q}^2, \ldots, \mathbf{q}^K\}$
Begin
Obtain $\mathbf{O} = \{\mathbf{o}^1, \mathbf{o}^2, \ldots, \mathbf{o}^K\}$ for Q;
while($t < t^{max}$) do
for($k \leftarrow 1$ to K) do
Perform feature selection according to $\mathbf{o}k$;
Train the SVM in the training set with the selected features;
Perform querying and compute $fit(\mathbf{o}^k)$ with the testing set;
end
Find the chromosome index $k_{opt}$with the highest querying accuracy;
for($k \leftarrow 1$ to K *and* $k \neq k_{opt}$) do
Perform mutation for $\mathbf{q}^k$;
Obtain $\mathbf{o}k$ for $\mathbf{q}^k$;
end
for($k \leftarrow 1$ to K *and* ($r = rand(0, 1)) \leq Pc$) do
Perform crossover operation;
end
$t \leftarrow t + 1$;
end
Find $\mathbf{o}_{opt}$with the highest querying accuracy;
Output subsets;
End

Fuzzy support vector machine querying mechanism

Fuzzy support vector machine is employed in this research work in order to decrease the training time and also to improve the efficiency of the querying mechanism. It is dealt in several literatures that the last step of pre-processing is scaling the training data by which normalizing all features which results in 0 mean and a standard deviation of 1 [12]. The FSVM is implemented with the idea that the input data are nonlinearly mapped into a high dimension feature space where a linear separating hyperplane is created to separate the two-group data in order to find an optimal separating hyperplane with maximum margin. Radial Basis Function (RBF) kernel function is employed in the FSVM. The fuzzy logic mechanism which is employed will avoid the numerical instabilities during the SVM calculation. Subsequently the important features are extracted in terms of the values of the parameters θj, the prudent fuzzy rules are applied based on the support vectors which lies as l=1 and N_s discovered by the SVM.

The training process is performed as follows:

1) Each support vector corresponds to a fuzzy rule. The number of fuzzy rules equals to the number of support vectors;

2) Given the ith support vector x_s^i ; i=1,…,L

a) The premise part of the ith fuzzy rule is evaluated as follows: the MF of fuzzy set for the jth input variable in the ith rule is

$$A_i^j(x_j) = a^j \quad (7)$$

Where m_i^j is the jth element of the ith support vector x_s^i.

b) The consequent part of the ith fuzzy rule is induced from α_0 and query labels, i.e., the consequent value of the ith rule is

$$b_i = \alpha_0^{(i)} y_s^{(i)} \quad (8)$$

where$\alpha_0^{(i)}$represents non-zero $\alpha_0^{(i)}$ and $y_s^{(i)}$ is the query label corresponding to the ith support vector x_s^i .Th query I membership function of x is defined using the minimum operator for

$$m_i(x) = \min_{j=1.n.} m_{ij}(x) \quad (9)$$

If x is satisfied

$$nD_k(x)\begin{cases} > 0 \; for k = i \\ \leq 0 \; for k \neq i, k = 1,..n \end{cases} \quad (10)$$

The fuzzy rule selection procedure is described by the following steps.

Step 1: Evaluate the misqueryification rates (MRs) of the rules on the validation dataset and the test dataset separately.
Step 2: Set s=1 and assign a small value to threshold $h_s(h_s > 0)$
Step 3: Select the most influential fuzzy rules by

$$\left\{Rule_i \middle| \alpha_0^{(i)} or w_i > h_s\right\} \quad (11)$$

Step 4: Construct a fuzzy query (FQ) by using the influential fuzzy rules selected in Step 3.
Step 5: Apply FQ to the validation dataset v and the test dataset t to obtain new MRs: Ev(s).
Step 6: If Ev(s) =Ev(0) , stop the selection and use FQ(s-1) as the final compact query and Et(s-1)as the measure of generalization performance for FQ(s-1); Otherwise, increase s by 1, assign a higher value to threshold h_s, and go to Step 3.

RESULTS AND DISCUSSIONS

Epilepsy patient's image dataset has been obtained from [20]. 75 patient image data is obtained. When the epilepsy case(s) image is queried the proposed AGA-FSVM mechanism retrieves the appropriate images from the image dataset with less time delay. Figure-1 shows the sample image dataset. Figure-2 shows the accuracy of the proposed AGA-FSVM mechanism.

From the Figure-2 it is evident that the proposed AGA-FSVM provides better accuracy. Time taken for query response is also analyzed and is shown in Figure-2.

From the result it is shown that the proposed AGA-FSVM takes less time to respond for the query.

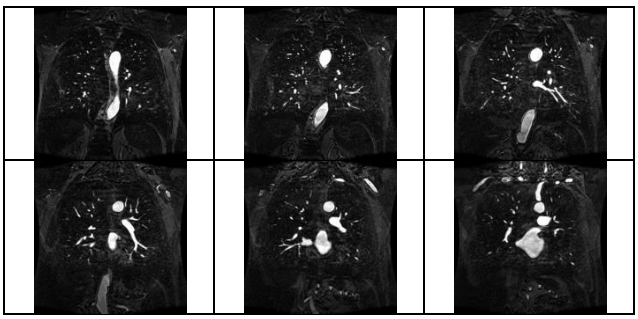

Figure-1. Sample image dataset.

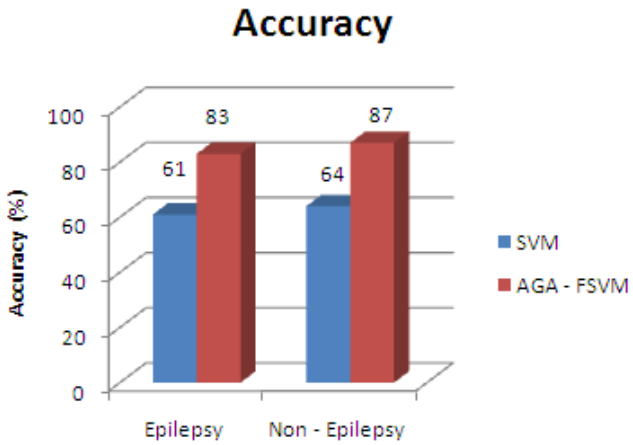

Figure-2. Accuracy.

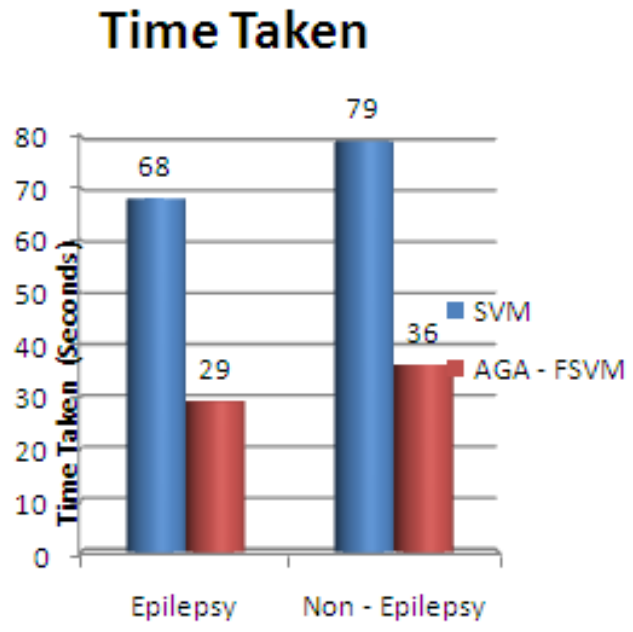

Figure-3. Time taken.

CONCLUSIONS

This research work presented Adaptive Genetic Algorithm based Fuzzy Support Vector Machine (AGA-FSVM) Query Mechanism for Image Mining among large image datasets. Adaptive Genetic Algorithm is modelled with five strategies in order to perform feature selection. FSVM is employed for querying the image dataset. The proposed work is evaluated using the performance metrics accuracy and time taken for query retrieving. The proposed AGA – FSVM is compared with SVM and the results demonstrated that the AGA – FSVM attains better results.

REFERENCES

[1] A. Jain, D. Zongker. 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19: 153-158.

[2] I. Guyon, A. Elisseeff. 2003. An introduction to feature and variable selection. Journal of Machine Learning Research. 3: 1157-1182.

[3] X.Liu, A.Mondry. 2005. An entropy based gene selection method for cancer classification using microarray data. BMC Bioinformatics. 6: 76.

[4] Z. Xu, I. King, M.-T. Lyu, R. Jin. 2010. Discriminative semi-supervised feature selection via manifold regularization. IEEE Transactions on Neural Networks. 21: 1033-1047.

[5] M. Dash, K. Choi, P. Scheuermann, H. Liu. Feature selection for clustering-a filter solution, in: Proc. Second Int'l Conf. Data Mining. pp. 115-122.

[6] R. Caruana, D. Freitag. Greedy attribute selection, in: Proc. 11th Int'l Conf. Machine Learning. pp. 28-36.

[7] J. Zhang, H. Deng. 2007. Gene selection for classification of microarray data based on bayes error, BMC Bioinformatics. 8: 370.

[8] H. Peng, P. Long, C. Ding. 005. Feature selection based on mutual information criteria of max dependency, max-relevance and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27: 1226-1238.

[9] C. Ding, H. Peng. Minimum redundancy feature selection from microarray gene expression data, in: Proc. Second IEEE Computational Systems Bioinformatics Conference. pp. 523-528.

[10] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek. 2010. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 32: 1582-1596.

[11] Y. D. Chun, N. C. Kim, I. H. Jang. 2008. Content-based image retrieval using multire solution color and texture features. IEEE Transactions on Multimedia. 10: 1073-1084.

[12] L. Nanni, J. Shi, S. Brahnam, A. Lumini. 2010. Protein classification using texture descriptors extracted from the protein backbone image. Journal of Theoretical Biology. 24: 1024-1032.

[13] M. Carlin. 2001. Measuring the performance of shape similarity retrieval methods. Computer Vision and Image Understanding. 84: 44-61.

[14] Text of ISO/IEC 15 938-3 Multimedia Content Description Interface- Part 3: Visual. Final Committee Draft, ISO/IEC/JTC1/SC29/ WG11, 2001. Doc. N4062.

[15] M. Swain, D. Ballard. 1991. Color indexing, International Journal of Computer Vision. 7: 11-32.

[16] L. Cieplinski. Mpeg-7 color descriptors and their applications, in: Proc. of 9th International Conference on Computer Analysis of Images and Patterns Seville. pp. 11-20.

[17] M.Stricker, M.Orengo. Similarity of color images, in: Proc. SPIE Storage and Retrieval for Image and Video Databases. pp. 381-392.

[18] J. Smith, S. Chang, Single color extraction and image query, in: IEEE International Conference on Image Processing. 3: 528-531.

[19] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, W. Niblack. 1995. Efficient color histogram indexing for quadratic form distance functions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 17: 729-736.

[20] Website: https://physionet.org/physiobank/database/images/.