Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/16658
DOI URL: http://dx.doi.org/10.21474/IJAR01/16658

RESEARCH ARTICLE

INNOVATIVE MODEL TO AUGMENT SMALL DATASETS FOR CLASSIFICATION

Dr. Vandna Bhalla
Delhi University, Department of Electronics, Sri Aurobindo College, Malviya Nagar, New Delhi, 110017.

Manuscript Info

Manuscript History
Received: 10 February 2023
Final Accepted: 14 March 2023
Published: April 2023

Key words:-
Augment, Immune System, Clonal Selection, Small Size Datasets, Bag-of-Words, Affinity, Avidity, Mutation and Crossover

Abstract

A vast number of applications do not have ample training data and consequently the accuracies suffer. There is a need to develop a technique to optimally and intelligently augment such datasets and design an automated classifier which gives good performance despite the smallness of dataset size. Meticulously stored data will ease retrieval. Artificial Immune System (AIS) is one amongst many computational algorithms in literature that are inspired by the dynamic learning mechanism of the human system. AIS based classification algorithm was proposed initially as one of the machine learning techniques which is suited for supervised learning problems. There are various applications areas of this powerful algorithm. We present a novel technique inspired by clonal selection to augment small datasets and classification.

Introduction:-

Small Data is crowding our personal devices and it needs a different approach from Big Data. Recent years have seen a growth in learnings related to small datasets.Even a small and limited number of synthetic samples can boost a dataset favourably provided the data augmentation approach is skilful. A study by Chang [1] puts forth the learning characteristics of small sized datasets. Pinto inhis work [2] enumerates the limitations of small size training sets. Optimal Augmentationof dataset is a significant challenge. Artificial data has been used often to increase the datasetsize. These techniques do not add authentic new data but the artificial and/or synthetic datadoes improve the outputs and makes it feasible to train with less data. The concept of virtualdatageneration to enhance recognition potential was initially put forth by Niyogi et al [3].

He created virtual samples by using previous information from a given small sized dataset. The scheduling complication in FMS (flexible manufacturing systems) is solved using the MTD (mega trend diffusion) techniques by Li et al [4] and they too used virtual samples to enhance the number of training examples for their research. For a data set as low as five, their classification precision improved from 69.3% to 78.23%. The virtual data generation does enhance the data size but under the perpetual risk of data favouritism which can result in over populating certain classes. Figure 1 lists the core issues faced while dealing with small data sets and known popular approaches and techniques to address them.

Corresponding Author:- Dr. Vandna Bhalla
Address:- Delhi University, Department of Electronics, Sri Aurobindo College, Malviya Nagar, New Delhi, 110017.

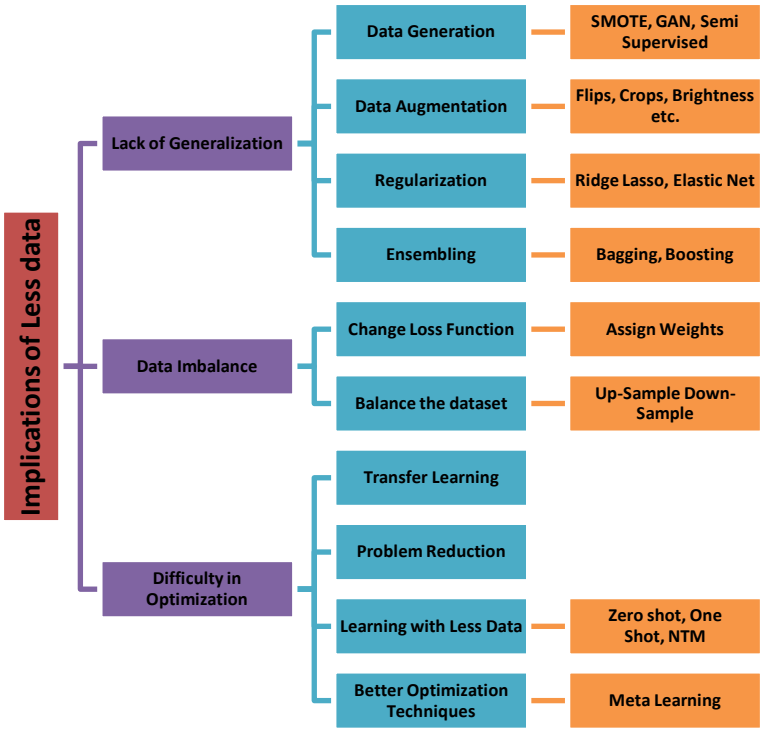

Figure1:- Basic implications of fewer data and possible approaches and techniques to solve it.

Our proposedArchitecture for classification with feature selection after augmenting data set is motivated by the human immune systems [5]. The intelligent novel data management system is based on Clonal Selection process of the human immune system [6]. The novel framework forms clusters for a dataset with inbuilt feature selection [7] [8]. The system creates a set ofrepresentative memory cells (Feature Vectors) using smart evolutionary mechanism which can then be utilized for the task of automatic classification to facilitate fast, accurate and easy retrieval. The developed model has the following features:

1. An inbuilt capability for pattern recognition of input data.
2. It captures the inherent diversity which maybe present within a class in the data.
3. It optimally augments training data inspired by clonal selection principles.
4. It performs classification and has the ability to perform well despite the smallness of ofthe training data size.
5. It is self-learning.
6. It is self-evolving.

The novel model returned encouraging results when tested on a collection of personalphotos belonging to three different classes.This framework works well especially for personal data collections.The training repositories of such data are very small and diverse. The evolving hybrid model isadaptive and assists users to organize their huge and diverse data piles. Compared with othercontemporary frameworks the proposed model gives encouraging results. TheExperiments were conducted taking different categories of hand-crafted features which are explained in detail. The results are reported at the end followed by conclusions.

Framework of the Proposed Model

Our immune system has antibodies and these capture foreign elements called antigens. If the specific antibody is not present in the system to combat the particular antigenthenmutation creates the required antibody and these subsequently proliferateand many more are generated [9]. This auto adaptive and self-learning mechanism of the human immune system inspires our model [10], [11],[12]. The model is divided into five phases: -

1. Representation of Data i.e the Memory Cells.
2. AIS inspired CS for Intelligent Data Augmentation.
3. Choosing the Appropriate Similarity measure.
4. Training the Model.
5. Testing the Model

Representation of Memory Cells

Each input is represented by its feature vector of dimension (m). Feature vectors are labelled as the memory cells or antibodies in our framework. For M images in the training phase there will be M memory cells. The initial size of memory set or the antibody set is therefore (Mxm). Each memory cell/feature vector in the memory set is chosen as sample antigen and clonal selection is performed based on its affinity with the rest of the memory cells.The top n memory cells each of dimension m generate clones based on affinity and avidity. Avidity represents the functional affinity and is the total strength of assorted affinities of an antigen with the antibodies. The number of clones for each memory cell is calculated as: -

$(Num_{clonei}) = \gamma \times A(n_i;G)$

For all $i \in N$. Here $A(n_i;G)$ represents the affinity of the ith memory cell/antibody with the antigen. So at the end of this step total number of clones are:-

$$total_{clonei} = \sum_{i=1} (Num_{clonei})$$

And $\gamma$is the constant which controls the size of all new generated antibody. The number of clones are directly proportional to affinity. Each set of these clones undergo mutation.

Mutation Simulation α 1/Affinity

This further populates the memory with the mutated clones. These are yet again checked for their affinity with the sample antigen. If the new memory cells satisfy the criteria, they are added to the antibody set else the process is repeated till the desired number of memory cell are obtained. This process is repeated for all memory cell in initial population of memory set (M). After Completion of this step,

the number of cells in the antibody set = Initial + new added

= (M+βM). Where β can be 1.....n.

The antibody set = (M+βM) x m

If there are 2 classes of memory size (P x m) and (Q x m) then the resultant antibody set will be: -

(P+βP) x m and

(Q+βQ) x m

The threshold for each class is calculated. Let class/antibody set consist of M feature vector/antibodies.

For all antibody $\in class_i$

do

{

- The affinity measure of chosen antibody with the rest of the memory set and store the value in the local array.
- Choose the maximum affinity ones from the local-array and store in the final-array.

}

set Class-threshold= Max/Min/Median(final-array)

After completion of this step, the model has all its classes optimally populated with representative feature vectors.

Training

The feature vectors of all training images make theinitial population of each class, N.Training Algorithm is as follows:

1. Initialize randomly (N), an antibody classand (M), the memory cell class.
2. A pattern is selected from the population and its affinity, A, is established with each cell of N and highest n are kept for further process.
3. Using the principle of Clonal selection,the clones aregenerated, and the total number is calculated by
$$C_{num} = \eta * A(I_1; I_2)$$
4. The new clones so generated are augmented with the initial population.
5. Repeat with all individual patterns in initial set.

The training generates a large pool of antibodies and for each distinct class. The clonal selection thus helps in generation of new intelligent data using existing data as depicted in Figure 2.

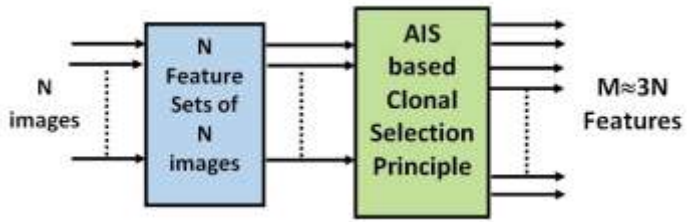

Figure 2:- Data Augmentation using Clonal Selection.

Testing

This phase ascertains the category of a testing input data. We have designed and implementeda two-layer classifier, akin to the immune system. The model uses the count and affinity combined values to categorize data. All data are represented by an appropriate representation i.e their feature vectors. All the categories are populated by the representative patterns(antibodies). The testing pattern (antigen) in Layer 1 is matched against the antibodies of each class. When matched with the particular class, all the antibody which qualify the threshold criteria are counted and if the count is above a threshold the class is chosen to contest for Layer 2 testing.

For Layer 2 testing a combined score of Count(the number of best match antibodies) andAvidity(the accrued strength of all affinities) determines the final allocation of the class. The Process of the Count and Affinity is shown in Figure 3. The example shown has three classes. The test data is represented in blue colour. The antibody in each class that qualify the threshold criteria are shown in the respective fields. Four for class 1, five for class 2 and three for class 3 qualify the threshold criteria. The closer the antibody to the test image the better is the affinity. The combined value (Count + Affinity) is used as the score. The class with the highest score is chosen as the class of test data.

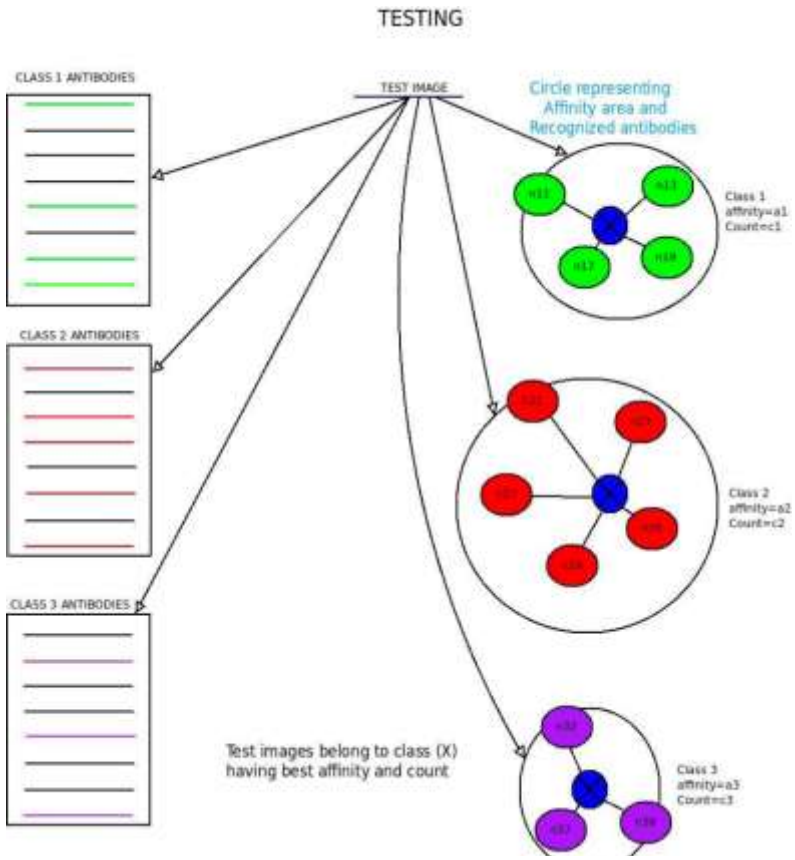

Figure 3:- Count and Affinity Concept.

Testing with Bag of VisualWords

The model was tested using Bag of Visual Words (BOVW) which is the sparse representation ofimage by a vector of the frequency count of the vocabulary of the local image features [13],[14]. It isbasically the sparse HISTOGRAM that is representing how frequently the visual words occurand each image is represented by the count of occurrence of the visual words, Figure 4. [15],[16]

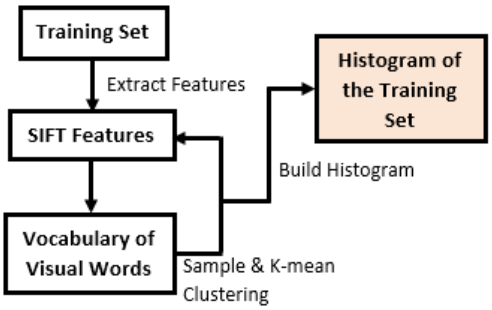

Figure 4:- Representation of image using BOVW.

The process of implementing the BOVW consists of two steps: -

1. Local feature detection and extraction: SIFT algorithm is used for detecting and extracting the local feature. SIFT feature vectors are invariant to any scale, translation or the rotation of the object. Here the scale space The key points with four parameters (the center coordinates X and Y, Radius, Angle) are obtained and their descriptors are calculated. Each detected region is represented with the SIFT descriptor with the most common parameter configuration with the orientation in 8 direction and block of size 4x4, resulting in a descriptor of 128 dimensions. An image of size n xm is represented with the N key points, each of dimension 128.

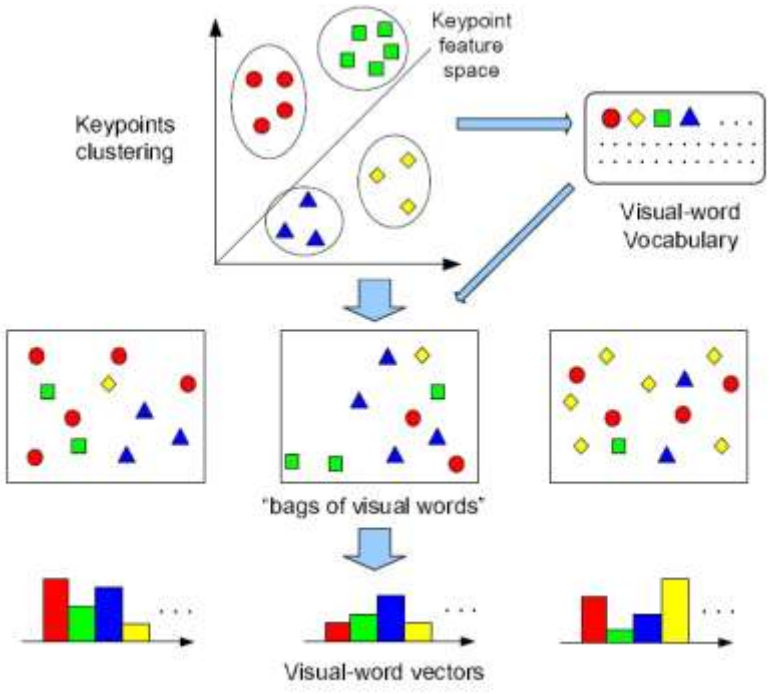

Figure 5: - Example showing Codebook creation.

2. Codebook Construction [17]: SIFT features suffer from the problem of varying cardinality and meaningless order. To overcome this VECTOR QUANTISATION technique is employed where we cluster the key point descriptors into the feature space using the Kmeans clustering algorithm. Each key point is encoded by the index of its cluster. All the key descriptors now possess their index number. Total number of key points mapped

to a cluster are counted. Hence a vector of size equal to the number of the clusters is obtained representing the histogram of visual words. The bag of visual words (BOVW) signifies the local pattern characteristics of the image. Figure 4.6 explains the process. The second row are images which are represented by n number of SIFT feature descriptors. These descriptors are mapped in the feature space having vocabulary size of 4. The descriptors are clustered using the K-mean algorithm and result in the histogram correspondingto each image as shown in row 3 of Figure 5.

Tuning the parameters is critical as the settings affect the accuracies to a great extent. During implementation we found the following parameters relevant: -

1. **Vocabulary Size**: The vocabulary size determines the size of feature vectors. On heuristic basis depending on the application, various cluster size was explored. As the cluster size increases the discrimination power of the features increase while the generality decreases. Smaller vocabulary size causes the dissimilar point to map to the same cluster leading to less discriminative features. There is a clear trade off between the generality and discrimination power and optimization is required.
2. **Normalization**: Performance of the classifier increases on normalizing the feature vector. The vector across the images is normalized. There are many visual words of image. Suppose the visual word t_i appears c_i^j times in the image X_i, then

$$t_i = \frac{c_i^j}{\sum_j c_i^j}$$

where $i \in [1, K]$, $j \in [1,N]$ and N is the number of images.

Histogram Intersection and Euclidean Distance, both were studied as similarity measures toascertain affinities.

S. No.	Class	Euclidean Distance			Histogram Intersection		
		Co	Mutation	No AIS	Co	Mutation	No AIS
1.	Picnic	55.2	51.8	49.5	65.6	57.4	50.1
2.	Wedding	57.6	50	49	68.5	60.1	51
3.	Conference	56.5	50.3	49.2	66.7	58	50.5

Figure 6:- BOVW Results on Small Dataset Images.

Conclusion:-

We performed our experiments on various parameter setting and the best are reported. As seenfrom Figure 6 crossover (Co) [18] is preferred over mutation and Histogram Intersection [19] seems abetter similarity measure option. All values are in percentage. Moreover, superior classificationaccuracy is achieved with AIS than without AIS. This endorses the power of AIS as the relevantdata generator as well as a classifier.The BOVW treats images as a document and devised its own technique for identifying "words" in the images. It follows three steps to achieve this i.e feature detection, its description,then followed by codebook generation. Its a kind of histogram representation of the features.One of the major limitation of this technique is that it completely overlooks the spatial relationshipsbetween the patches which are very significant while classifying photos. A feature extraction that can capture the co-occurrence of the spatial features will yield better results.

References:-

1. Fengming M. Chang, Characteristics analysis for small data set learning and the comparisonof classification methods, Proceedings of the 7th WSEAS International Conferenceon Artificial Intelligence, Knowledge Engineering and Data Bases (Stevens Point,Wisconsin, USA), AIKED'08, World Scientific and Engineering Academy and Society(WSEAS), 2008, pp. 122–127.
2. James J DiCarlo, Nicolas Pinto, and David Daniel Cox, Why is real-world visual objectrecognitionhard?, (2008).

3. Partha Niyogi, Federico Girosi, and Tomaso Poggio, Incorporating prior informationin machine learning by creating virtual examples, Proceedings of the IEEE 86 (1998),no. 11, 2196–2209.
4. Der-Chiang Li, Chih-Sen Wu, Tung-I Tsai, and Yao-San Lina, Using mega-trenddiffusionand artificial samples in small data set learning for early flexible manufacturingsystem scheduling knowledge, Computers & Operations Research 34 (2007), no. 4,966–982.
5. Vandna Bhalla, Artificial Immune System: Intelligent Systems inspired by the human Immune System, Vol 05 N0. 01 Jan – Mar, 2023 of IJEMMASSS, ISSN: 2581-9925
6. AndrewWatkins, Artificial immune recognition system (airs): Revisions and refinements, Genetic Programming and Evolvable Machines, 2002, pp. 173–181.
7. R. Kheddam and A. Belhadj-Aissa, Classification of remotely sensed images using clonal selection theory of artificial immune system, 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA), Oct 2014, pp. 1–6.
8. Emma Hart and Jonathan Timmis, Application areas of ais: The past, the present and the future, pp. 483–497, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
9. L Nunes De Castro and Fernando J VonZuben, The clonal selection algorithm with engineering applications, Proceedings of GECCO, vol. 2000, 2000, pp. 36–39.
10. Revathi M and Arthi K, Application of Artificial Immune System Algorithms in Dataset Classification , International Journal of Innovative Research in Advanced Engineering (IJIRAE) 1 (2014).
11. Grzegorz Dudek, An artificial immune system for classification with local feature selection., IEEE Trans. Evolutionary Computation 16 (2012), no. 6, 847–860.
12. Jieqiong Zheng, Yunfang Chen, and Wei Zhang, A survey of applications, Artificial Intelligence Review 34 (2010), no. 1, 19–34.
13. Gidaris, Spyros, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. "Learning representations by predicting bags of visual words." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6928-6938. 2020.
14. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., & Pérez, P. (2021). Obow: Online bag-of-visual-words generation for self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6830-6840).
15. Sun, Huadong, Xu Zhang, Xiaowei Han, XuesongJin, and Zhijie Zhao. "Commodity image classification based on improved bag-of-visual-words model." Complexity 2021 (2021): 1-10.
16. Arun, K. S., V. K. Govindan, and SD Madhu Kumar. "Enhanced bag of visual words representations for content based image retrieval: a comparative study." Artificial Intelligence Review 53 (2020): 1615-1653.
17. Saini, Manisha, and Seba Susan. "Bag-of-Visual-Words codebook generation using deep features for effective classification of imbalanced multi-class image datasets." Multimedia Tools and Applications 80 (2021): 20821-20847.
18. Hassanat, Ahmad, Khalid Almohammadi, Esra'A. Alkafaween, EmanAbunawas, AwniHammouri, and VB Surya Prasath. "Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach." Information 10, no. 12 (2019): 390.
19. Chakravarti, Rishav, and Xiannong Meng. "A study of color histogram based image retrieval." In 2009 Sixth International Conference on Information Technology: New Generations, pp. 1323-1328. IEEE, 2009.