International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.6, No.7, July 2017 DOI:10.15693/ijaist/2017.v6i11.8-14

Development and Investigation of Novel Genetic Based Ensemble Data Stream Classification Model

Dr.S.Jayanthi.
Professor,
Department of Computer Science and Engineering,
Samskruti College of Engg and Technology,
Ghatkesar – 501 301, Hyderabad, India.
nigilakash@gmail.com

Abstract - Data stream classification is an inevitable task in digital sectors for the past decade. This research work proposes a novel Genetic based Ensemble Data Stream Classifier to confront the plights arises in the data stream classification process. The proposed classifier has been formulated with the ensemble of neural networks, fuzzy logic, K-Means and genetic algorithm.

The investigation of this research theoretically justifies and empirically ascertains the compatibility and competency of the proposed ensemble in the data streaming environment. The efficacy of the proposed model is tested by conducting exhaustive experiments with a dedicated server providing video services to the users and has been compared with the state of the art data stream classification algorithms with respect to various parameters.

Key words: Fuzzy Logic, Genetic Ensemble Classifier, K-Means Classifier, Data Stream Classifier

I. MOTIVATIONS

This research work expounds a novel genetic based ensemble data stream classifier which intended to synergize both the supervised and unsupervised classification approaches so as to enrich the data stream classification task. In this research work, Neuro Fuzzy classifier and Genetic K-Means have been chosen as the instance of supervised and unsupervised classifier respectively. The classifier hybridization in the proposed ensemble has not been explored in any other related research works in the literature. However, a plenty of research works have attempted to resolve these problems; most of them have been investigated with synthetic data streams where the attributes of classes and distribution center

are not concealed. Hence, it is apparent that the algorithm which works efficiently on synthetic data streams may not be good for online data streaming environment where the attributes and features of a class generated by different distribution centers are obfuscated. Hearkening to this, the proposed work is deployed in online to assert its efficacy over various evaluation metrics with other contemporary algorithms.

II. RELATED WORK

In this section, the key features of comparative algorithms are discussed for empirical investigation. Most recent data stream classification methods, Online Accuracy Updated Ensemble (OAUE), Mine Class Method (MCM), and Enhanced Classifier for Data Streams with novel class Miner (ECSMiner) are taken to comparing the results.

A. Online Accuracy Updated Ensemble

OAUE algorithm has been formulated to combat against concept drifts in the data stream classification task by combining both the key principles of incremental and block based learning algorithms. It adopts windowing strategy to train and evaluate its component classifiers. Its limitation is that its accuracy level fluctuates between different kinds of the data streams (Brzezinski *et al*., 2014).

B. Mine Class Method

MCM algorithm is also aimed at resolving concept drifts in data stream mining process. It employs Gini Coefficient, K-means clustering and Adaptive Threshold techniques to confront various issues of data stream classification. It is inferred from the literature is that MCM actively identifies novel classes, but is neither dynamic and nor robust in resolving concept drifts (Masud *et al*., 2014).

C. Enhanced Classifier for Data Streams with Novel Class Miner

ECSMiner is an ensemble based data stream classification algorithm aimed at discovering novel classes, in the presence of concept drifts. This algorithm adopts the decision tree and K-Nearest

Neighbour algorithm as component classifiers along with the silhouette coefficient to discover novel classes. It is observed that its accuracy level deteriorates when the concept drift rate is high (Masud *et al*., 2013).

III. ROLE OF BASE CLASSIFIERS IN THE ENSEMBLE

This section emphasizes about the suitability and functionality of the base classifiers chosen for the proposed classifier, Genetic based Ensemble Data Stream Classifier. The ensemble of Fuzzy Neural Networks and Genetic K-Means algorithm has been formulated to execute the data stream classification process in a real-time environment.

A. Neural Networks

Neural networks are prominent classifiers that are capable of making human-like cognitive decisions in prediction and classification process through the experience gained on the classification model. Neural networks have neurons as the basic construction units which process the input data and get fired when it entails significant stimuli (Han and Kamber, 2006).

In precise, neurons combine the input data along with its weight, which either magnify or dampen the role of input data in the classification task. Products of the input data weight are summed and then passed to the activation function which determines whether and to what extent that the input data affect the classification task. A sample neuron in a neural network is depicted in Fig.1. However, the neural networks have several layers, namely, input layer, hidden layer and output layer to carry out classification task.

The features of neural network such as adaptability, parallel processing, robustness, and scalability make them easy to learn on data streams. Moreover, if neural network size is fixed, there is no more memory management problem arises during the classification task. This feature makes them fit for data stream classification. For the investigation of this research, the Back Propagation algorithm (BPA) which employs multilayer feed forward neural network to carry out the classification, has been chosen due to its high efficacy and robustness in handling unknown data.

This neural network holds one input layer to receive the input parameters, a number of hidden layers to process the input parameters and one output layer to produce class labels where each unit in all the layers are initially assigned with weights. Number of output units in the output layer is subject to the number of class labels defined in the classification algorithm. The class labels are determined by repeatedly updating the weights of the neurons in the hidden layers which is further sent to output layer for producing class labels.

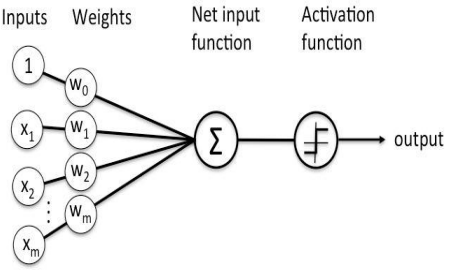

Fig.1. Neuron in a neural network

Initially BP algorithm gain knowledge on processing the training data set. Then the neural network set with the right number of parameters and hidden layers are trained to predict the class labels. The classification results of back propagation algorithm and that of neural network are compared repeatedly so as to ultimately reduce the mean square error of the classification result and thereby improve the accuracy of the classification process. Mean square error is computed by using the following formula:

Mean Square Error (MSE) = _____

Here y'_i is the predicted value of the variable y_i and d is the number of tuples in the test data set.

When the MSE rate is higher than the preset threshold value, then the BPA algorithm updates the weights of the neurons in the multilayer feed forward neural network in backward direction from the output layer towards input layers. It is also inferred that the number of input parameters and hidden layers significantly influence the classification result. Hence, they need to be set intuitively with suitable number of parameters and layers.

B. Fuzzy Logic

Neural Networks work well on classifying raw data with high learning abilities. However, fuzzy logic has explanation abilities with reasoning on a higher level using linguistic information acquired from domain experts, but they cannot learn or adjust themselves to a new environment. Fuzzy Inference Engine is depicted in Fig.2. (Chen 1989).

In the fuzzy inference engine process, input data are fed to fuzzification interface which has the ability to interpret the linguistic attributes. The processed input data are further sent to the fuzzy inference engine which makes inferences on the arrived data with fuzzy rule base which consists of well-defined rules for each possible class.

For example, let the truth values for A and B are 0.3 and 0.6 respectively. Inference rules R1 and R2 can

be formulated by defining the minimum and maximum values as given below.

R1: A ∧ B=min (0.3, 0.6) = 0.3

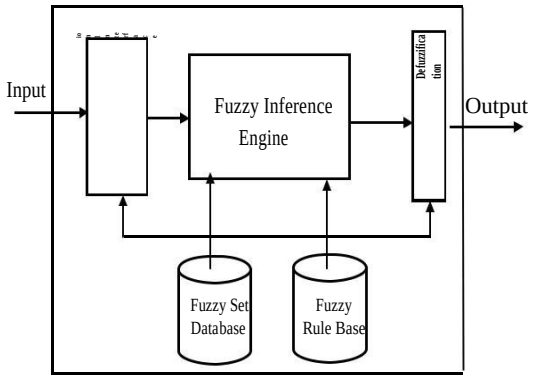

Fig.2. Fuzzy Inference Engine

In addition with fuzzy rule base, a fuzzy set database is also available in the fuzzy inference engine which helps to correlate with the historical data in the inference engine.

Being attracted to the key features of neural networks and fuzzy logic and the congeniality between them, it is contrived to deploy its synergy as one of the modules in the proposed system. That is, fuzzy networks have been formulated to achieve the learning abilities and high computational power of neural networks and explanation abilities of fuzzy systems. Furthermore, When a representative set of instances are available, a neural fuzzy system can automatically transform it into a set of fuzzy IF-THEN rules, and thereby reduce the reliance on expert knowledge when designing classification modules.

Neuro-Fuzzy classification module in the proposed system works up with 16 attributes of the Fuzzy Set, F, where

F = {Do It Yourself, Drama, Sport, Movie, Funny, Technology, 2D-Animation, 3D-Animation, Entertainment, Tutorials, Short Film and Vector}

A membership function value of 0.5 for a genre specifies that the video attributes do not indicate any existing video genre, while the membership function value of 1.0 points out a very high likelihood of a particular video genre, and a member function value of 0 indicates a very high likelihood that the attributes do not belong to a particular video genre.

C. K-Means

K-Means is a prominent unsupervised classification algorithm in which the number evolving clusters and its labels are not preset and known only at the time of classification. This algorithm clusters the instances of data streams by computing the distance between the instances of data streams. In precise, the algorithm is initially consigned with the number of clusters and the data set of instances.

Initially, the algorithm is set with number of clusters and a dataset containing instances of data streams. Then it chooses a cluster center for each cluster randomly. In subsequent phases, the algorithm computes the mean value of each cluster based on the initial instances and relocates the instances to other clusters so that they can be located to the nearest cluster center. This process is repeated until all the instances are located to the nearest cluster center.

The Algorithm for K-Means clustering is given below:

Algorithm: K-Means clustering
Input: K: the number of clusters, D: a data set containing instances
Output: A set of clusters

1. Select k instances (seeds) chosen from D as initial cluster centers;
2. Assign each instance into the nearest cluster by calculating the mean value of the instances in the cluster;
3. Update the mean value of the instances in the cluster;
4. Repeat Steps 2 and 3 until no change. Nearest cluster can be identified by computing the distance between the instances using the Euclidean distance.

Its drawback is that the cluster centers are fluctuated with respect to the randomly chosen seeds or instances. This algorithm suffers much when the data set entails noise. Also, this algorithm does not guarantee to provide a global optimum solution.

D. Genetic Algorithm

The genetic algorithm is an algorithm which is widely adopted for optimizing the objective function. To achieve this, the algorithm executes selection, crossover, and mutation operations over the instances of the population along with fitness value. Hence, to optimize the objective function, mean square error, of the k-means algorithm, genetic algorithm can be suitably adopted. However the synergization of k-means and genetic algorithm is investigated in several research works, it has not been carried out in an online real-time environment to classify the data streams (Jayanthi et al, 2014).

E. Modified Genetic K-Means Algorithm

To circumvent the sensitivity of K-Means to initial points and to achieve the global optimum solutions that improve the clustering performance in real-time environment, it is contrived to employ

modified genetic k-means algorithm to classify unlabeled instances of upcoming data streams

.

Procedure for Genetic K-Means algorithm
Input: Initial population, Di= (x1, x2,--.,Xn), minimum and maximum clusters (Kmin, Kmax);
Output: Cluster centers, C;

1. Begin
2. For data chunk Di ∈S do
3. Set MaxGen to max iteration allowed Gen =1
4. begin
5. Generate a number N, between Kmin and Kmax
6. Choose K_i points (rows) randomly from Di
7. Distribute these data points randomly in the chromosome
8. Set unfilled positions to null value
9. End
10. For Gen=Gen+1 do
11. Calculate Fitness value for each chromosome in the population
12. Begin
13. Extract the K_i cluster centers stored in it
14. Perform clustering by grouping each point in the cluster to the closest center
15. Calculate DB index (DBi) by Euclidean distance
16. Compute fitness as $1/DB_i$
17. End
18. If Gen<Maxgen
19. Select Single point and crossover with the probability value NB
20. Mutation performed with the probability value NB
21. Begin
22. Randomly choose one position of chromosomes.
23. If this position is null, choose a point from Di randomly and make it as a cluster center, C
24. Else set this position to null
25. End
26. Use the detected C for k-means algorithm
27. End

In the above algorithm, minimum and maximum numbers of clusters are set to 1 and 10 respectively. The minimum number of clusters is set to 1 with the assumption that there might be at least one novel cluster may emerge at any time upon the arrival of unlabeled instances.

The maximum number of clusters is set to 10 with an assumption that there will not be more than 10 novel classes may emerge at the same time in the online environment. However, the maximum number of clusters is subject to modification if required.

IV. THE PROPOSED GENETIC BASED ENSEMBLE DATA STREAM CLASSIFIER

The proposed system, Genetic based Ensemble Data Stream Classification (GEDSC), is centered on investigating the coherence and competency of Neuro-Fuzzy classifier and Genetic K-Means classifier in accomplishing data stream classification. However fuzzy logic is prominent for interpreting instances and rules, it is passive on generalizing the classification rules, and in contrast, neural network is active on generalization but not so on interpreting the classification rules. Neuro-Fuzzy, the ensemble of neural network and fuzzy logic has been formulated to achieve the complementary strength of both the algorithms for data stream classification.

K-Means algorithm is a prominent distance based unsupervised classifier that classifies instances by calculating the interclass and intraclass similarity which is often getting convoluted in fixing the floating central tendency of clusters. Genetic algorithm has been prudently chosen to tune up the central tendency of K-Means algorithm with its unique operations, namely, selection, crossover and mutation.

Data streams escalated from clients are scanned and sliced into data chunks using the sliding window technique. Genetic Based Ensemble Data Stream Classifier employs two major phases on the data streams as given below.

- In the first phase, Neuro-Fuzzy classifier classifies the labeled instances of data chunks under normal condition, in case of no concept drift.
- In case of any concept drift and concept evolution, Genetic K-Means algorithm is applied to enact on data stream classification. In precise, Genetic K-Means algorithm isolates the instances of data streams which are aberrant and tracks to find inter class similarity between them and with the upcoming data to assert the existence of novel classes and concept drifts.

The steps taken in the proposed system are given below:

1. Initially, the data streams are sliced into fixed sized data chunks so as to cope with the infinite length of data streams.
2. The sliced data chunks are buffered and scanned by the fixed sized sliding window so as to perform preprocessing at the preliminary level. However the size of the sliding window is adjustable, for ease of handling of the data chunks, it is kept fixed.
3. If all the instances in the data chunk are labeled, then can be classified with fuzzy neural network.

Else repeat the process from the step 4 to step 9 repeatedly until all the instances are classified with some class labels.

4. If any concept drift or concept evolution is found with the instances of the data chunks, Modified Genetic K-Means algorithm is employed on classifying the instances.
5. Modified Genetic K-Means algorithm finds interclass similarity between the instances of the data chunks with the maximum number of cluster centers.
6. If the interclass similarity is high with respect to any cluster center of the maximum number of clusters, it is declared as a novel class. Else repeat the process from step 7 to step 8 to determine whether the instances formulate novel classes or outliers.
7. The instances are tracked for some time, so as to find similarity between the upcoming instances of the data streams.
8. If the similarity is found high, then the instances will be declared as a novel class, else will be discarded as outliers.
9. The declared novel classes are fed back to the fuzzy neural networks to recognize the classes of the instances of the upcoming data chunks.

V. EXPERIMENTAL SETUP

A. Hardware and Software Specifications

To implement the proposed system, Enterprise STD dedicated server having the configuration IBM 160 G8 Series, 1 X Intel Hexa Core Xeon Processor E5-2620, 2.0 GHz, 15 M Cache, 7.2 GT/s Intel QPI is used for 6 month period to run our video stream mining applications. This model is tested by generating video requests from 25 personal computers having the configuration, Windows 8.1 OS, I5 processor and Visual Studio-13.

However, the request for videos can be generated from any type of the devices such as mobile, laptop, etc. A video emulator, Internet Streaming Video for Video Service Provider for Infinite Data Streams (ISVPFin) has been configured to play videos on the client side.

Moreover, the investigation carried out with 25 systems shall not assert the supremacy of the online server. Hence, a drift emulator has been designed to generate a large number of requests for various types of videos. The user interest on videos is collected by a real time dedicated server to prebuffer the anticipated videos in advance so as to reduce the video playout time, viewing latency and play out distortion. The real time server considers 16 attributes of the user demands and the corresponding 12 classes, namely, Do It Yourself, Drama, Sport, Movie, Funny, Technology, 2D-Animation, 3D-Animation, Entertainment, Tutorials, Short Film, and Vector.

B. Drift Emulator

When there is a significant change found among 18 attributes such as, protocol, user name, IP address, source data header, destination data header, login status, geographical class, login status, number of failed logins, service count, video type, total data size, duration, resolution, bit rate, repeat status, number of audio channels, audio quality and category, it is considered as concept drift.

Notable challenges in achieving the task of video service providing to a user are twofold:

- Firstly, the user interest on videos is dynamic which leads to concept drift and
- Secondly, the release of a new category of videos, which might subsequently captivate the interest of users in viewing these videos, is also very common which might lead to concept evolution.

If the user is not active for more than 30 days, then the user is considered as an inactive user and will be removed from the classifier to speed up the learning process so as to service the active users in a better way. Performance evaluation of this empirical investigation is carried out to assert the competency of the proposed classifier over other comparative algorithms:

Through the exhaustive empirical investigations, it is observed that the performance of the proposed method is the most competent of all the comparative algorithms. The average performance report shown in Table.1. is generated by tracking the efficacy off all the algorithms considered for the investigation over various sizes of data chunks in online mode.

C. Results and Discussions

From the empirical results, it is proven that the proposed Genetic based Ensemble Data Stream Classifier is the most competent of all the other comparative algorithms in terms of accuracy, precision, recall and F1-Measure. It is also evident that the proposed classifier is the most reliable to carry out the data stream classification task on all varying sizes of data chunks. In the graphical plots, memory consumption is measured in KiloByte (KB) units. The response time and processing time are measured in microseconds.

Furthermore, it is also inferred that the method ECM-BDF achieves high memory utilization, but its accuracy rate is not up to the level of the proposed classification model. Similarly, the method OAUE is good at response time compared to other algorithms, but its accuracy rate is also not up to the level of the proposed model.

Table 1 Average performance report of GEDSC and other comparative algorithms

Procedure	Accuracy	Precision	Recall	F1-Measure	Memory	Processing Time	Response Time
OAUE	84.38	84.308	83.6	83.95	6116224	**1966**	**790**
MCM	84.50	84.431	83.83	84.13	6124217	2063	887
ECS MINER	84.41	84.41	83.72	84.06	6146313	2076	901
ECM-BDF	85.02	85.007	84.48	84.74	**6100636**	2005	831
CDRTREE	86.30	86.319	86.32	86.32	6494881	2211	1056
GEDSC	**87.60**	**87.83**	**87.65**	**87.74**	6434304	2189	1025

It is apparent that the proposed classifier's efficacy is negligibly lower in terms of memory utilization and response time than other comparative models and this might be due to the inherent features of the Back Propagation neural network algorithm, which is used as a component classifier in the Neuro-Fuzzy classifier, whose efficacy varies with respect to the number of hidden layers and the weights initially assigned to its connection units.

Hence, it is suggested that the algorithm chosen for ensemble shall be revised to strengthen the memory utilization and response time. The classifiers efficacy can also be evaluated by varying the number of hidden layers and weights of its connection units in the back propagation neural network. In addition, it is also planned to deploy an increased number of diverse contemporary algorithms to pep up the efficacy of the proposed classifier in all aspects.

VI. CONCLUSION

Significant aspects of this research work are delineated underneath:

- Resolving concept drift and concept evolution are the major thrust in data stream classification.
- The proposed classifier, Novel Genetic based Ensemble Data Stream Classifier is intended to execute data stream classification task by resolving its key issues such as concept drift and concept evolution.
- The proposed classifier has been deployed in a real-time dedicated server to provide the anticipated videos to the viewers by knowing their interest.
- Viewers' interest in viewing videos may change over time and leads to a concept drifting scenario in the data stream classification process. The server is consigned to recognize the changing interests of the users over videos.
- New kind videos may emerge at any time, which shall captivate the users' interest in viewing these videos. This scenario leads to concept evolution issue and the server is consigned to recognize the changing interests of the users on trendy videos and provide service accordingly.
- Fuzzy neural network and modified genetic algorithm are designed intuitively and investigated exhaustively.
- Fuzzy neural network takes a key role in classifying labeled instances of data streams in the absence of concept drift and concept evolution.
- Modified genetic k-means algorithm takes a key role in addressing concept drift and concept evolution.
- Genetic algorithm is chosen to optimize the cluster tendency issue of the K-means algorithm.
- Here Back Propagation Neural Network has been chosen as the representative classification model of neural network, which has the ability to learn the massive amount of data.
- The extensive empirical investigation ascertains the competency of the proposed model in term of accuracy, precision, recall and F1-Measure.
- However, the classification model deteriorates negligibly on response time and memory utilization over some comparative algorithms.
- The limitation observed in the proposed model might be due to the inherent feature of BPA, which entails prolonged time for the classification task.

REFERENCES

1. Dariusz Brzezinski and Jerzy Stefanowski, "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm", IEEE Transactions on Neural Networks and Learning Systems, 25(1): 2014, pp.81-94.
2. Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, "Ashok Srivastava and Nikunj C. Oza, Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams", IEEE Transactions On

Knowledge And Data Engineering, 25(7), 2014, pp.1484-1497.

3. Mohammad M. Masud, Member, Latifur Khan, Jiawei Han and Bhavani Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", IEEE Transactions on Knowledge And Data Engineering, 23(6), 2011, pp.859-874.
4. Jiawei Han and Kamber, "Data Mining Concepts and Techniques", Margan Kaufman Publisher, Second Edition, 2006.
5. Chen. Y.-Y. "The Global Analysis of Fuzzy Dynamical Systems. University of California", Berkeley, 1989.
6. Liu Jing, Xu Guo-sheng, Zheng Shi-hui, Xiao Da and Gu Li-ze, "Data streams classification with ensemble model based on decision-feedback", The Journal of China Universities of Posts and Telecommunications, 21(1), 2014, pp.79–85.
7. Jayanthi, S., B. Karthikeyan, "A Recap on Data Stream Classification", Adv. in Nat. Appl. Sci., 8(17), 2014, pp.76-82.
8. Jayanthi, S., B. Karthikeyan, "Aggregate eighted Ensemble Model for Data Stream Classification", International Journal of Applied Engineering Research (IJAER), 9(21), pp: 4945-4949.
9. Jayanthi, S., B. Karthikeyan, "Incremental Aggregation Model For Data Stream Classification", ARPN Journal of Engineering and Applied Sciences, 10(8), 2015, pp.3828-3832.
10. Sobolewski P and Wozniak M., "Concept drift detection and model selection with simulated recurrence and ensembles of statistical detectors", Journal of Universal Computer Science, 4 (19), 2013, pp.462-483.

Author Profile

Dr.S.Jayanthi was born in the year 1981. She received her Bachelor degree in Computer Science from Bharathidasan University, India in 2002, and Master degrees in Computer Applications, and Computer Science and Engineering from Bharathidasan University, in 2005, and from Anna University of Technology, India in 2009, respectively. She obtained her Ph.D from Karpagam University, Coimbatore, India. She has 8 years of teaching experience. Her area of research includes Data Mining, Neural Networks and Big Data analysis. So far she has published 26 papers in various national, International conferences and journals.