Discovering CQA post voting prediction using Artificial Neural network and Entropy Analysis

Kavita Shinde, Sarita Patil

Abstract*: As the whole world is aware of education and its importance the knowledge in web pages also grows due to crowd sourcing. Many web portals are running because of the good amount of the investment of the user's knowledge for free and in a well desired manner makes the portals make hefty business. Some web portals like stack overflow, yahoo and even some social media sites like twitter and all are completely relying on crowdsourcing data. Most of the time it is hard to identify the best answer from the users for a question that was raised by the other user in the portal. Some methodologies are existed to achieve this where they are using the scores that are given by the other users or likes. This many times yield in loss of precision and never cross check the validation of the answers with their contents. So this paper puts forwards an idea of identifying the bag of word technique along with the Artificial neural network and entropy analysis of for nonlinear and unplanned distribution of data. Finally, by using the Bayesian law along with the fuzzy classification model for predicting degree yields the best prediction of question and answers.

Index Terms: CQA, ANN, Bayesian Probability, Entropy Evaluation, Fuzzy Logic, Bag of words.

I. INTRODUCTION

Most of the community question answering system is using the likes and dislikes and direct integer scores to predict the best question answering. This proposed research article uses artificial neural network and Shannon information gain entropy theory to achieve the excellence in the prediction of the question and answers. Some of the entities are introduced in this section to their core strengths and working patterns which are extensively used in the proposed model.

K means - K-means is the oldest and, most frequently used unsupervised learning algorithm for clustering. It is model depend clustering method describing the model in centroids terms which is taken as the mean of all the points and is applied to objects in an exceedingly continuous n-dimensional area

The k-means algorithmic program involves indiscriminately choosing k initial centroids wherever k could be a user outlined a range of desired clusters. Every single point is then appointed to a nearest centroids and the collection of point near to a centroids from a cluster. The centroids get updated in step with the points within the cluster and this method continues till the points stop swapping their clusters. K-means is sometimes run again and again, beginning with completely different random centroids every time. The results are compared by examining the clusters or by a numeric measure like the clusters' distortion, that is the add of the square variations between every data point and its respective centroids. The clustering with the lowest distortion rate is selected as the best clustering in the cluster distortioncase. Entropy evaluation - A new technique to evaluate entropy is to check the sample of random sequence expected entropy with the computed entropy of the sample. This can be the primary known technique that takes into examining the size of the sample sequence and its impact on the accuracy of the computation of entropy. In different types of engineering application such as genetic analysis, independent component analysis, image analysis, manifold learning and speech recognition for evaluation of time delay system and biological system status the differential entropy is estimated to find out some observations. The histogram based estimation is the most easiest and general approach for entropy evaluation. For selecting a technique for entropy evaluation the differentiating factor that are considered are the information distribution nature, trade off amongst variance of the estimates and bias.

ANN - ANN (Artificial Neural Network) framework depends on the functions and structure of the biological neural network. Data that flow among the network affects the formation of the ANN because a neural network changes depend on the input and output. It is taken as nonlinear statistical information modeling tools where the complicated relation between input and output patterns are found. ANN is also known as a neural network. Advantages of ANN, but the most important one is that this network, learn from the observing data sets. Theses tools are very cost effective and ideal technique for arriving at solutions. ANN considers data samples only rather than whole data sets for finding solutions. This feature of ANN saves both money and time. There are three layers in the ANN which are interconnected. The main objective of the ANN is to resolve the problem in the same manner as the human brain does. Various tasks in which ANN is used such as machine translation, computer vision, speech recognition, video games, medical diagnosis, etc.

Fuzzy Classification - Fuzzy classification is the technique for organizing or grouping the elements into fuzzy sets, whose associate function is explained by fuzzy propositional

function truth value. In another way fuzzy classification is stated as fuzzy set of individuals having same characteristic. A fuzzy propositional function is similar to an expression having more than one variable, in such a way when values are declared to each variable, the expression becomes a fuzzy proposition. It is also applied to geographic objects. It also applied for analysis of vegetation,

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

soil composition and other situation which changes gradually in physical composition. There are various fuzzy classification methods such as K nearest neighbors, Fuzzy C-means etc.

In this paper, section 2 is dedicated for literature review of past works. Section 3 describes the proposed methodology and Section 4 discusses the results and evaluation of the proposed technique. Finally Section 5 concludes this paper with future extension possibilities.

II. LITERATURE REVIEW

This section of literature survey eventually reveals some facts based on thought analysis of many authors work as follows.

GengZhang, Han-Xiong Li, [1], proposed the idea of PFC (Probabilistic Fuzzy Classifier). They implant the theory of probabilistic fuzzy to create PFC for the pattern categorization under fuzzy and stochastic uncertainties. The PFC performance is enhanced in comparison with pure probabilistic or traditional fuzzy method, by perfectly designing the probabilistic fuzzy inference and the secondary probability density function and by including probabilistic voting method. The probabilistic fuzzy rules derive from processed data or expert knowledge is helping to make decision more accurate and simple to understand. The probabilistic property encapsulated in the data can be referred as decision confidence level.

Keeley Crockett, Annabel Latham, David Mclean, Zuhair Bandar, James O'Shea [2] proposed a method to develop a fuzzy predictive model by utilizing fuzzy classification trees. They use their method to find out the different learning style of students through a conversational tutoring structure. They perform their experiments by using 41 independent variables dataset. All the experiments were carried out on two learning styles, understanding and perception. The results of the experiments prove that the fuzzy predictive model increased the accuracy rate of OSCAR CITS in comparison with a single predictor variable used within the system.

Jianping Gou, Wenmo Qiu, Qirong Mao, Yongzhao Zhan, Xiangjun Shen and Yunbo Rao [3] proposed the MLMNN (Multi-Local Based Nearest Neighbor Classifier). They first search all the k categorical nearest neighbors from a query sample and then utilize to calculate k categorical multi-local mean vectors. The k-categorical multi local mean vectors represent a distinct local class-specific sample distribution. The linear combination of k categorical local mean vector and each local mean vector representation coefficient linear combination is used to represent each query sample. The k categorical multi local mean vectors and query sample reconstruction residual is used to find the each query sample class label.

Prithwish Jana, Soulib Ghosh, Suman Kumar Bera and Ram Sarkar [4] proposed k-means clustering method for degraded normalized image. They used the dataset of H-DIBCO 16 which contains highly depraved documents images which are handwritten and calculate each image detailed result. The computed result is then compared with the top three winners in the contest. Their technique shows the best result in performance.

Jie Sun, Zhi-Min Liu [5] proposed k means clustering algorithm to decrease the interference between UEs in distinct clusters. A resource allocation technique is also utilized to lower the co-channel interference between neighbor users. They compare their method with location depend grouping technique and simulation result prove that their algorithm performance is better in reducing the interference amongst UEs.

G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu and B. Kingsbury [6] present a computerized system for quality monitoring of recorded calls in a call center. They integrate the pattern matching, speech recognition and maximum entropy technique to measure call quality and give rank according to the call quality. At both end calls are tagged as interesting. They used the pattern matching technique to answer standard quality control question. The system is used to improve human monitoring.

Giuseppe Bianchi, Chiara Carusi, Lorenzo Bracciale [7] proposed both online joint inference and convenient model to sort out the problem of a correct answer to a set of labeling tasks or binary choice questions in a crowdsourcing system. To find out the solution of problem like worker's abilities to answer questions and question difficulty level they first cast Bayesian framework which allow obtaining closed form single question inference step. After that they address more general framework. Finally they, proposed the computational efficient algorithm which increases the entropy reduction for the assigned questions.

Maorong Shao, Ying Zhang, Ying Jiang, Lingxuan Zhu [8] proposed a system for the refugee crisis. They used the dynamic programming model and entropy AHP for their system. The main influencing factor is found by applying Analytic Hierarchy Process. The author used entropy weight and proposed Entropy AHP Math Model for evaluating refugee crisis. Principal Component Analysis Technique is used to evaluate the environmental factor that changes over time. The logistic model is used to predict the number of refugees.

Driss El Hannach, Rabia Marghoubi, Mohamed Dahchour [9] presents a PPMIS prioritization approach. The PPMIS approach applied information entropy technique for prioritizing and final selecting PPMIS of all possible. The approach involves users and PPMIS decision taker and keep their judgment under uncertainty.

Ji Zhang, Zhi Du, Dong Xie, Shouxia Jiang, Yang Liu, Jin Ma and Yanbo Chen [10] proposed uncertainty evaluation model to resolve the difficulty of an existing design scheme of substance. In the proposed model, each index weight is found by entropy weighted fuzzy comprehensive evaluation technique and then applying Monte Carlo simulation technique to simulate the uncertain factors in the new generation smart substation design scheme.

Samuel Jonathan Slade [11] introduce the problem related to neural network approach and prove that the stagnation detection impact, on autopoiesis of the system and self organization.

Utku Kose [12] introduced a software system based on artificial neural networks to improve the educational processes. The present system evaluates the performance of each student in terms of their multiple intelligence levels and given them more effective course contents to improve their learning. The author has also done some evaluation works to find how effective their system is in improving student performance.

H. M. Peixoto, A. A. R. Diniz, N. C. Almeida, J. D. de Melo, A. D. Dória Neto, A. M. G. Guerreiro [13] presents a system for tracking moving individual or object. The different method like artificial neural network, digital image processing, and reinforcement learning with TD (temporal difference learning) is used to develop this system.

H. M. Peixoto, A. A. R. Diniz, N. C. Almeida, J. D. de Melo, A. D. Dória Neto, A. M. G. Guerreiro [14] proposed Bayesian Artificial Immune System (BAIS), to architect group on neural network for classification. Both selection and generation of components are made by using BAIS. A probabilistic model of combined allocation is built by BAIS to generate new candidate solution. The probabilistic model of Bayesian network is used to capture automatically expressive variables, interaction and then identifying and found partial solutions to the problem.

Pablo A. D. Castro and Fernando J. Von Zuben [15] proposed AIS (Artificial Immune System) for studying feed forward ANN topologies. The Artificial Immune System inspired algorithm enhanced each node activation function and explore the space of neural network topologies. The author used 7 classification model datasets to prove the effectiveness of the proposed system for designing accurate and effective neural network classifiers.

Toby O'Hara, Larry Bull [16] presents a learning classifier system having lookahead neural network. The lookahead neural network in the present system decreases the generality of the rules. The author highlights the nature of supervised learning of the anticipatory task and amends each system rule with ANN.

III. PROPOSED SYSTEM

The proposed methodology of the community question answer prediction system can be deeply narrate based on the below mentioned steps as shown in the figure 1.

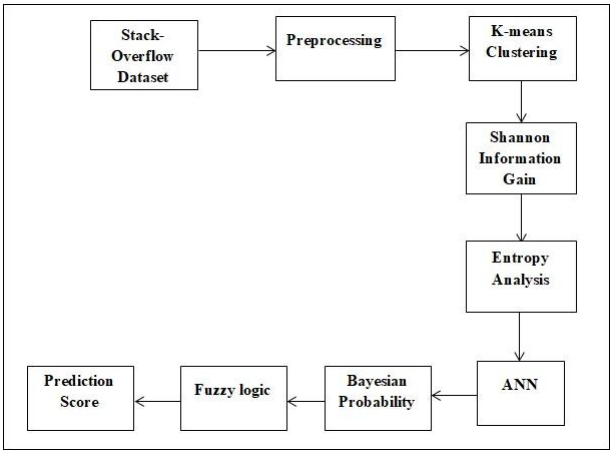

Figure 1: System overview of the proposed System

Step 1: Data Collection, preprocessing and labeling - This is the initial step of the proposed model where model collects the all the attributes from the dataset or database like Owner UserId, Creation Date ,user ID, Parent ID, Score, and body of the posts etc. into a double dimension list. Then this list is being subjected to attribute selection process where only required attributes like Parent ID, Score and body of the posts are being selected to form a new double dimension list of string data.

Once the attributes are selected, then attribute " body of the posts" is selected for the preprocessing process as this is the only one attribute that is selected in the form of a string. To Preprocess the string series of series of steps are taken as narrate below

Tokenization - Here each of the words in the post string is fetched in a list by using the split function of Java strings. Then these words are embedded into a list for the further processing.

Special Symbol removal - Each of the embedded words are retrieved from the list to replace the special symbols with the empty strings to get rid of the same.

Stopword Removal - As we know natural narration always contains abundant of unwanted words which always creates a problem to get the text into a valid and good structure of numerical value. So natural language processing or machine learning techniques is used to convert the unorganized text into some numerical values.

As the first step towards this proposed model identifies the stopwords or conjunction words in the English language which are statically stored in a list. Once they are identified, then they are replaced with the empty string to get rid of them. This makes the text light weight and also helps to retain the original meaning of the text.

Stemming - After the removal of conjunction words, there is one more final work is remained as to bring down the word in its base form. This process again makes the text more lighter along with the retaining of core semantic.

Labeling - Once the complete post text is preprocessed then it is subjected to convert into a numerical value based on the term weight . This is done according to the bag of words that eventually indicates the positivity regarding the post.

Step 2 : K means Clustering - This is the step where clusters are formed for the labeled data list, Where the data of the row contains parent ID, Score and then the feature weight. Initially this list is subject to evaluate the mean Euclidean distance between the each row for the attribute like score and feature data by considering the point on x- axis. This can be shown in equation 1.

Then this estimated Euclidean distance is appended at the end of the respective row of each list. The segregated mean of these row Euclidean distances yields the average Euclidean distance of the whole list data itself. After this, random data points are selected based on the normalization factors of the list length percentage. Once these data points are estimated then the ranges of the Euclidean distances of these data points produces the estimated centroids ranges . Which can be shown with the below mentioned equation 2 and 3.

Then finally each row is

added to the respective ranges according to their Euclidean distances which ultimately yields the clusters in a proper manner.

$$Ed = \sqrt{(x1 - x2)2 + (y1 - y2)2} \quad \text{_____(1)}$$

$$min_b = D_{pi} - E_d \quad \text{_____(2)}$$

$$max_b = D_{pi} + E_d \quad \text{_____(3)}$$

Ed = Euclidean Distance
D_{pi} = Data Point
min_b = Minimum Boundary Range
max_b = Maximum Boundary Range

Step 3- Entroy Evaluation - This is the step where most important PIDs are evaluated based on the entropy. Here all the PIDs are enlisted into two lists, then from one list unique elements are extracted. Then, based on these unique PID's each of the PID's distributions is estimated in the other PID list. Then this is applied in shannon information gain theory to get the most distributed PIDs as mentioned in the below equation 4. Shannon information gain theory yields the entropy value in between 0 and 1 for each of the PIDs. Then these PIDs are sorted in descending order so that the upper half of the list is considered as the most distributed and eligible PIDs.

$$IG(E) = -(P/T)\log(P/T) - (N/T)\log(N/T) \quad \text{___(4)}$$

Where
P= Frequency of the PIDs present count in the original list
N= Non presence count
T= Total number of PIDs in the data.
IG (E) = Information Gain for the given PID

Step 4: ANN -Bayesian theory -This is the most important part of the proposed model, where by using the clusters formed in the past step using K-means algorithm are used to identify the best among them. This is done by evaluating the count of the most distributed PIDs which are extracted using the Information gain theory in the past step. Then the cluster with the largest count is considered as the best fit neuron for the further process.

Once the best neuron is selected, then it is subject to deep layer evaluation where the mean and standard deviations are evaluated for the Euclidean distances of this neuron as mentioned in the equation 5,6 and 7.

$$\mu = \frac{(\sum_{i=1}^{n} Edi)}{n} \quad \text{_____(4)}$$

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Edi - \mu)^2} \quad \text{_____(5)}$$

$$f(T_h) = \mu + \delta \quad \text{_____(6)}$$

Where
δ - Standard Deviation
μ- Mean
Edi - Euclidean distance of instance row
N- Number of Rows in cluster
$f(T_h)$ - Neuron threshold function

Once the mean and standard deviations are evaluated , then by using these a threshold value is being estimated through the equation 6. And then all the feature values of the rows are estimated for their value greater than the threshold and count them to fit bayesian law. This ratio eventually results as the peak probability of the prediction which is used by the fuzzy classification function to predict the community question answering posts.

Step 5: Fuzzy Classifications - Once the peak probability is evaluated, then it is subjected fuzzy classification process. Where a distance is evaluated among the minimum, i.e 0 and the peak probability factor and it is divided by 5. As the fuzzy classification theory contains the 5 probability factors like VERY LOW, LOW , MEDIUM , HIGH AND VERY HIGH.

Each of these division factors are segregated as in the form of ranges to get the fuzzy rules and thereby to classify the probability ratio of the PIDs into the prediction classification model based on the below mentioned fuzzy classification algorithm 1 .

ALGORITHM 1: FUZZY CLASSIFICATION

```
//Input : PID Score Vector Pv
//Output: Prediction Classified list CL
1: Start
2: Set small=0, big=0
2:    For i=0 to size of Pv
3:       TSet = Pvi   [ Tset = Temporary Set]
4:       Sc=Tset[1]
5:                IF ( Sc <small) [ sc= Score]
6:                  small=Sc
7:                  IF(Sc>big)
8:                  big=Sc
9:    End for
10:          d=( big-small)/5    [ d= Distance ]
11:             For i=1 to 5
12:                IF(i==0)
13:             Fc(min=small, max=d) [ Fc = Fuzzy Crisp Set
]
14:                else
15:                Fc(min=Fci-1(max),max= Fci-1(max)+d
16:             End For
17:                For i=0 to Size of Fc
18:                   For j=0 to size of Pv
19:                   TSet = Pvi   [ Tset = Temporary Set]
20:             Sc=Tset[1]
21:                IF Sc ∈ Fci
22:                add   TSet   to
CL
23:                END IF
```


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

24: End For
24: **End For**
25: **return** C_L

IV. RESULTS AND DISCUSSIONS

The proposed model is deployed in real time stand alone mode using the Java based windows machine. For the performance evaluation of the system, proposed methodology uses a machine of standard configuration with 4GB primary memory and Core i5 processor. System uses windows based java enabled machines with Netbeans 8.0 as IDE and MySQL 5.0.22 as Database Server.

Precision and recall are considered as the best performance measuring parameters. Precision can be defined as the ratio of number of relevant prediction detected to the sum of number of relevant and irrelevant prediction detected. Relative effectiveness of the system is well expressed by using precision parameters.

Whereas the recall can be defined as the ratio of number of relevant prediction detected to the sum of relevant prediction not detected. Absolute accuracy of the system is well narrated by using recall parameters.

Precision and recall can be more clearly elaborated as follows.

- X = the number of relevant prediction detected,
- Y = the number of relevant prediction not detected, and
- Z = The number of irrelevant prediction detected.

So, Precision = (X/ (X+ Z))*100
And Recall = (X/ (X+ Y))*100

No of Given posts	Relevant predictions identified (A)	Irrelevant Predictions Identified (B)	Relevant Prediction not identified (C)	Precision	Recall
25	20	2	5	90.90909091	80
50	40	3	10	93.02325581	80
75	68	10	7	87.17948718	90.66666667
100	84	8	16	91.30434783	84
125	107	14	18	88.42975207	85.6

Table 1: Performance Data through precision and recall

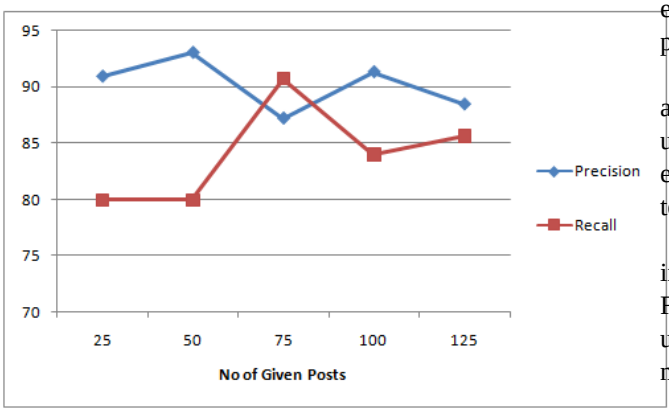

Figure 2: Performance Evaluation through precision and recall

On plotting the graph for precision and recall for different number of runs we found some facts that system yields 90.16 % of precision and 84.05 % of recall.

When Proposed model is compared with that of [17] which is Analyzing the sentiments based on the unsupervised and supervised Co- occurrence Data. Unsupervised method of this paper uses association rule mining to identify the sentiments through co-occurrence data. On the other hand supervised technique uses co-occurrence through weight matrix method. As we know the Association rule mining includes more number of candidate sets which eventually yields large unnecessary relations between the sentiments.

Whereas the proposed model of post voting prediction model uses the ANN and Entropy evaluation approach which yields better results than this. This can be seen with the below mentioned table and plot.

Methods	Precision	Recall
ANN- Entropy	90.16	84.05
Unsupervised and Supervised	84.4	83.1

Table 2: Comparison Table

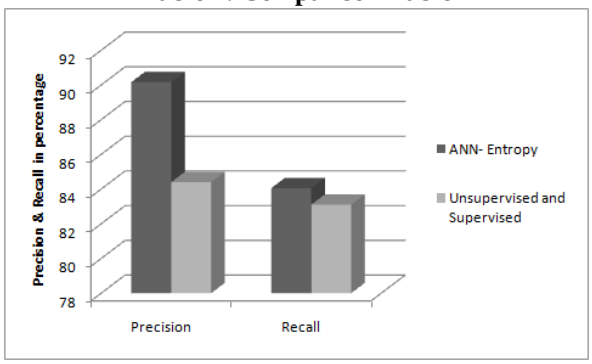

Figure 3: Comparision Evaluation of the Approach

V. CONCLUSION

The Proposed model of community question answering voting prediction system is deployed in real time stand-lone system with the features of stack overflow web portal. And the proposed model also uses the stack overflow data to evaluate the model as discussed in the past section.

The Model uses proper machine learning features to convert the string data of posts into numerical values. Then Shannon Information theory along with the K-means clustering generates a perfect distribution of PIDs. These evaluated PIDs play vital role in the identification of probability neurons using ANN and Bayesian model.

Proposed model predictions are classified using the fuzzy abstract classification. These predictions were evaluated using the precision and recall in the past section which eventually provides the best result on comparison of other techniques.

As the future work this proposed idea can be implemented in many differnt online community sites like Twitter , Facebook. This can be done to predict the live mood of the user using interactive web crawler in the huge deployment model using cloud computing.

REFERENCES

1. GengZhang, Han-Xiong Li, A Probabilistic Fuzzy Learning System for Pattern classification, DOI: 978-1-4244-6588-0/10, IEEE, 2010.
2. Keeley Crockett, Annabel Latham, David Mclean, Zuhair Bandar, James O'Shea, On Predicting Learning Styles in Conversational Intelligent Tutoring Systems using Fuzzy Classification Trees, DOI 978-1-4244-7317-5/11,IEEE,2011.
3. Jianping Gou, Wenmo Qiu, Qirong Mao, Yongzhao Zhan, Xiangjun Shen and Yunbo Rao, A Multi-Local Means Based Nearest Neighbor classifier, DOI 10.1109/ICTAI.2017.00075, IEEE, 2017.
4. Prithwish Jana, Soulib Ghosh, Suman Kumar Bera and Ram Sarkar, Handwritten Document Image Binarization: An Adaptive K-Means Based Approach, DOI: 978-1-5386-3745-6/17, 2017.
5. Jie Sun, Zhi-Min Liu, K-Means Clustering Algorithm for Full Duplex Communication, DOI 978-1-4673-9026-2/ 16, IEEE, 2016.
6. G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu and B. Kingsbury, 2d Color Barcodes For Mobile Phones Automated Quality Monitoring In The Call Center With ASR And Maximum Entropy, DOI: 1-4244-0469-X, IEEE, 2006.
7. Giuseppe Bianchi, Chiara Carusi, Lorenzo Bracciale, An online approach for joint task assignment and worker evaluation in crowd-sourcing, DOI: 978-1-5090-4260-9/17, IEEE, 2017.
8. Maorong Shao, Ying Zhang, Ying Jiang, Lingxuan Zhu, A Refugee Crisis System Based on Entropy AHP and Dynamic Programming, 978-1-5090-0729-5/16, IEEE, 2016.
9. Driss El Hannach, Rabia Marghoubi, Mohamed Dahchour, Project Portfolio Management Information Systems(PPMIS), DOI: 978-1-5090-0751-6/16,IEEE, 2016.
10. Ji Zhang, Zhi Du, Dong Xie, Shouxia Jiang, Yang Liu, Jin Ma and Yanbo Chen, Improved SEC Model Based Evaluation Approach for Design Scheme of The New Generation Smart Substation, DOI: 978-1-5090-5417-6/16, IEEE, 2016.
11. Samuel Jonathan Slade, A Reactively Learning Neural Network that Decides Behaviors for an Artificial Life System with Homogeneous Agents, DOI: 978-1-5090-4093-3/16, IEEE, 2016.
12. Utku Kose, An Artificial Neural Networks based Software System for Improved Learning Experience, DOI: 10.1109/ICMLA.2013.175, IEEE, 2013.
13. . Peixoto, A. A. R. Diniz, N. C. Almeida, J. D. de Melo, A. D. Dória Neto, A. M. G. Guerreiro, Modeling a System for Monitoring an Object Using Artificial Neural Networks and Reinforcement Learning,, DOI: 978-1-4244-9637-2/11, IEEE, 2011.
14. H. M. Peixoto, A. A. R. Diniz, N. C. Almeida, J. D. de Melo, A. D. Dória Neto, A. M. G. Guerreiro, Learning Ensembles of Neural Networks by Means of a Bayesian Artificial Immune System, DOI: 10.1109/TNN.2010.2096823, IEEE, 2010.
15. H. M. Peixoto, A. A. R. Diniz, N. C. Almeida, J. D. de Melo, A. D. Dória Neto, A. M. G. Guerreiro, Bayesian Learning of Neural Networks by Means of Artificial Immune Systems, DOI: 0-7803-9490-9/06, IEEE, 2006.
16. Toby O'Hara, Larry Bull, Building Anticipations in an Accuracy-based Learning Classifier System by use of an Artificial Neural Network, DOI: 0-7803-9363-5/05, IEEE, 2005.
17. Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker, " Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis With Co-Occurrence Data ", IEEE TRANSACTIONS ON CYBERNETICS,2017