American Journal of Applied Sciences 11 (5): 764-768, 2014
ISSN: 1546-9239

doi:10.3844/ajassp.2014.764.768 Published Online 11 (5) 2014 (http://www.thescipub.com/ajas.toc)

VIRTUAL MINING MODEL FOR CLASSIFYING TEXT USING UNSUPERVISED LEARNING

Koteeswaran, S., E. Kannan and P. Visu

Department of CSE, Vel Tech Dr. RR & Dr.SR Technical University, Chennai, Tamilnadu, India

Received 2012-10-07; Revised 2013-12-30; Accepted 2014-02-20

ABSTRACT

In real world data mining is emerging in various era, one of its most outstanding performance is held in various research such as Big data, multimedia mining, text mining etc. Each of the researcher proves their contribution with tremendous improvements in their proposal by means of mathematical representation. Empowering each problem with solutions are classified into mathematical and implementation models. The mathematical model relates to the straight forward rules and formulas that are related to the problem definition of particular field of domain. Whereas the implementation model derives some sort of knowledge from the real time decision making behaviour such as artificial intelligence and swarm intelligence and has a complex set of rules compared with the mathematical model. The implementation model mines and derives knowledge model from the collection of dataset and attributes. This knowledge is applied to the concerned problem definition. The objective of our work is to efficiently mine knowledge from the unstructured text documents. In order to mine textual documents, text mining is applied. The text mining is the sub-domain in data mining. In text mining, the proposed Virtual Mining Model (VMM) is defined for effective text clustering. This VMM involves the learning of conceptual terms; these terms are grouped in Significant Term List (STL). VMM model is appropriate combination of layer 1 arch with Analysis of Bilateral Intelligence (ABI). The frequent update of conceptual terms in the STL is more important for effective clustering. The result is shown, Artifial neural network based unsupervised learning algorithm is used for learning texual pattern in the Virtual Mining Model. For learning of such terminologies, this paper proposed Artificial Neural Network based learning algorithm.

Key words: Document Clustering, Virtual Mining Model, Unsupervised Learning, Analysis of Bilateral Intelligence

1. INTRODUCTION

Generally engineering problems are classified into two kinds, static and dynamic. The static natured problems are comparatively less complex because the behaviour of the problem is static till the end of its life time, whereas the dynamic natured problems are changing its parameters and attributes more frequently over time. Therefore the static natured problem requires comparatively less effort than the dynamic one. These problems are also analysed using mathematical models.

Usually solutions to dynamic natured problems in engineering discipline are classified into mathematical models and implementation model. The mathematical model comprises straight forward rules and formulae related to the problem definition of a particular field of domains. Whereas, the implementation model derives some sort of knowledge of the real time decision making behaviour such as artificial intelligence and swarm intelligence. It has a complex set of rules compared with the mathematical model. The mathematical model provides static rules which will solve problems in the most of the engineering field of study. But it has also some pitfalls (Zhong and Ghosh, 2003; Zhu, 2003; Kotsiantis and Pintelas 2004) in the optimality in the application to the dynamic natured problems. As the dynamic natured

Corresponding Author: Koteeswaran, S., Department of CSE, Vel Tech Dr. RR & Dr. SR Technical University, Chennai, Tamilnadu, India Tel: +91 9884378785

problem is changing its characteristics frequently over time, the solution to the problem also required few changes in the mathematical rule or entirely new set of rules.

The combination of k-means with spectral analysis (Zha *et al*., 2001), extended k-means algorithm (Kotsiantis and Pintelas, 2004), the Spatial Mining (Ng and Han, 1994) Principal Component Analysis (Jain *et al*., 1999) is a concept which based on the well-known image processing technique, the Locality Preserving Index (LPI) (Agrafiotis and Xu, 2002; Cai *et al*., 2005; 2011), Divide-and-merge (Cheng *et al*., 2006) conceptual model is proposed (Lebanon, 2006) in the literature which has performance limitations due to more epochs and repeated iterations.

In the proposed ANN based unsupervised learning, training data which contain text are data sets, improving accuracy of text clustering is the required output and achieving error free clustering is the goal. The architecture of the proposed ANN based unsupervised learning, training and testing methodologies, the sample data set and ratio of training and testing dataset are the important factors for achieving optimal result in a neural network based learning model. There is a variation of ANN model available, such as Neuron By Neuron (NBN), two Hidden Layers Artificial Neural Network (2HLANN).

The Neuron By Neuron (NBN) is a implementation, applied for nonlinear signal processor in the field of digital signal processing which is proposed by Wilamowski (2009). In this model, the traditional back propagation neural networks are improved. The proposed NBN is compared with Existing Error back Propagation algorithm (ERP). The ERP is the most powerful and popular learning model but it has few pitfalls, (1) slow processing which requires 100-1000 times more iteration and (2) less accuracy.

A 2 Hidden Layers Artificial Neural Network (2HLANN) model is proposed by Mkadem and Boumaiza, (2011). It is used for predicting the dynamic nonlinear characteristics of wideband power amplifiers. The 2HLANN is an improved model of feed forward neural network. The 2HLANN is designed in terms of number of neurons, learning rate and memory space.

2. MATERIALS AND METHODS

Text mining is one of the emerging research area in the domain of data mining. The data mining is an emerging technique which applies many approaches and methods from another field of study and also the data mining is implemented in another area to learn hidden knowledge (Koteeswaran *et al*., 2012a). In this proposed work, Artificial Neuran Network (ANN) based unsupervised learning is used for learning text in the Virtual Mining Model.

The proposed method provides efficient learning which identifies patterns which have synonymy and the convergent of the training algorithm is very fast than existing methodology. From the results, it is concluded that the performance of proposed Analysis of Bilateral Intelligence (ABI) is optimized. Hence, the proposed Virtual Mining Model with ABI learning will provide optimality than existing clustering algorithm.

The performances of algorithms and techniques employed in process field of domain area unit improved by means that of correct learning technique. Hence, so as to enhance the performances of operating VMM, a brand new learning technique is projected. This unsupervised learning technique is employed for classifying text within the Virtual Mining Model. It applies the training method to spot 2 equivalent terms (Bilateral) that has an equivalent that means. It contains text documents as datasets. Up accuracy of text cluster is that the needed output and its achieved error free cluster is that the goal.

The projected learning model involves the training of abstract terms from the VMM. The terms learned from the projected learning rule area unit classified and added to the STL. The frequent update of abstract terms within the STL is additional necessary for effective cluster (Koteeswaran *et al*., 2012b). For learning of such terminologies, this projected work applies Artificial Neural Network based mostly learning rule.

2.1. Learning and Testing using Unsupervised Learning Model

The ABI applies the learning process to identify two equivalent terms which have the same meaning (Koteeswaran *et al*, 2013). ABI contains text documents as datasets, improving accuracy of text clustering which is the required output and achieving error free clustering in a shorter time is the goal.

Here we proposed a working model which supports the layer 1 of the neural architecture with highness in the network model. Training and testing phase is done through the propagated model of ANN. Our model supports ABI with the outstanding performance in layer 1. The working model of ABI is forked as the parental model of our proposed model.

The working model of the proposed ABI Learning method is explained in the following sections.

The Activation Bias function which shown in Equation (2) is applied in the proposed ABI Equation (1):

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

where, X_A is the output in the hidden and output layer. Where the inputs are 'x_i' which is connected to the hidden layer from input layer. The connection has weights 'w_{ij}', between inputs to hidden layer and the output of the neurons referred as 'w_{jk}' is computational values between output and hidden layer. Where, 'b' neurons in the output layer, 'a' neurons in the hidden layer and 'i' neurons in the input layer. The detailed design diagram of neuron model is shown in **Fig. 1**.

Step 1: Input phase

The proposed ABI has implemented from well known initial phase. In input phase the data attribute is given as vector in the input layer. In the initial phase, the values of the weights are assigned. Let the values are 'w_{ij}' and 'w_{jk}'. Where the attribute vector value 'w_{ij}' is a value of the hidden layer and input layer. 'w_{jk}' is a value of output layer-hidden layer respectively. The other constants are penalty constant, which is defined as μ; and the number of iterations, which is called an epoch, is initialized in the system. The weight vectors 'w_{ij}' and 'w_{jk}' are to be optimized in order to minimize the error function.

Step 2: Bias Phase

The weight of the activation function is summed to train the neural model. Since the Bias function has high weight challenges, these weights were properly trained in the Bias activation phase with appropriate vector value. This weight adjustment step is processed based on sigmoid activation function, shown in the first phase.

Step 3: Optimization of Learning Layer:

$$S_{optimum} = A^{-1} x B \quad (2)$$

where, Equation (3 and 4):

$$A = \sum_{n=1}^{N} Z_a^n Z_i^n a, i = 1, ..., N \quad (3)$$

$$B = \sum_{n=1}^{N} Z_a^n t_b^n \; a, b = 1, ..., N \quad (4)$$

where, 'Z^n'= scalar output of the hidden neuron of training data 'n'. 'A' is the output of the hidden layer and 'B' is the output of the output layer. 'a' is the neuron in the hidden layer and 'b' is the neuron in the output layer, 'i' is neuron in the input layer and 't' is transaction function. The weights are trained in this phase. Where the actual weight function is given with appropriate weights.

Step 4: Test for completion

RMS error (E_{RMS}) was then calculated comparing the '$R^{test'}$ matrix with '$S^{optimum'}$ matrices calculated in Step 3 Equation (5).

a. $E_{RMS} < E$ (5)

The hidden layer weight matrix 'R' is updated 'R'= '$R^{test'}$. Decrease the influence of the penalty term by decreasing 'μ', Proceed to Step 5 Equation (6).

b. $E_{RMS} \geq E$ (6)

Increase the influence of 'μ' and repeat

Step 5: Output layer

In this the trained data sets are processed with high compatible bias function with initial activation vector. The results acquired in this section is seem to be predicted with undefined values

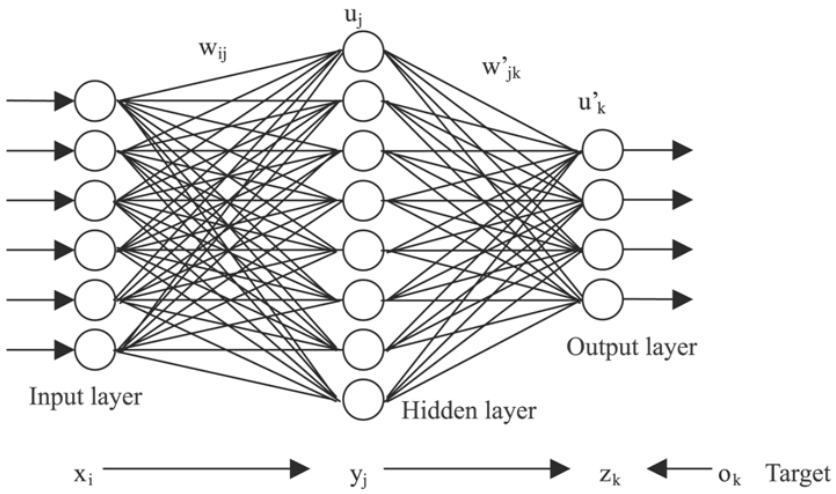

Fig. 1. Design of ANN model

Step 6: Process termination

If the RMS error is not within the desired range, repeat Step 3, else the training process is ceased. After the successful completion of the training phase, the sample real time data are given as input of the system. The system will choose comparatively best path. This thesis used 60% dataset for training and 40% dataset for testing.

3. RESULTS AND DISCUSSION

This ANN based learning model is implemented using Neural Network Tool Box in MatLab. In the training algorithm, the goal is assigned as "0.01" and the epoch is assigned as 250.

Figure 2 shows the % RMS error in the Estimation and Elimination of NBN, 2HLANN and the proposed learning model.

The estimation error identifies a number of documents and our terms identified in the clustering model. The elimination error defines the mismatch ratio for document clustering.

Comparisons of RMS error in estimation and elimination for proposed ABI learning model Vs existing models is shown in **Fig. 2** and also %error in estimation and elimination for proposed ABI learning model Vs existing models are shown in **Fig. 3**.

The result is shown in **Table 1** and performance is shown in **Fig. 2**, it is concluded that the performance of proposed ABI learning model always performs better than the existing methodology.

Figure 2 shows the proposed ABI learns the synonymy better than the existing systems. From this, it is concluded that the proposed ABI performs better than existing systems. The ABI shows around 30% improvement in the estimation and around 23% improvement in the elimination.

Evaluation of bias function in activation bias is described clearly in the **Table 1** which denotes the performance of the proposed model. RMS error rate and its proper analysing is done clearly in this model. So that our model combined with ABI outperforms with highness in performance and classification.

The convergence of the proposed ABI and existing learning models are compared in **Fig. 2**. The percentage RMS error in estimation is reached at 7.83% in NBN, 7.23% in 2HLANN whereas; it is only 4.60% in the proposed learning model.

The percentage RMS error in elimination is reached at 5.15% in NBN, 8.65% in 2HLANNwhereas; it is only 4.75% in the proposed learning model.

Table 1. Comparison of error growth on proposed model Vs existing models

No. of epoch	NBN	2HLANN	Proposed ABI	Activation bias
50	0.175	0.150	0.130	0.0012
100	0.140	0.110	0.080	0.0080
150	0.100	0.080	0.050	0.0050
200	0.075	0.060	0.025	0.0010
250	0.040	0.025	0.010	0.0010

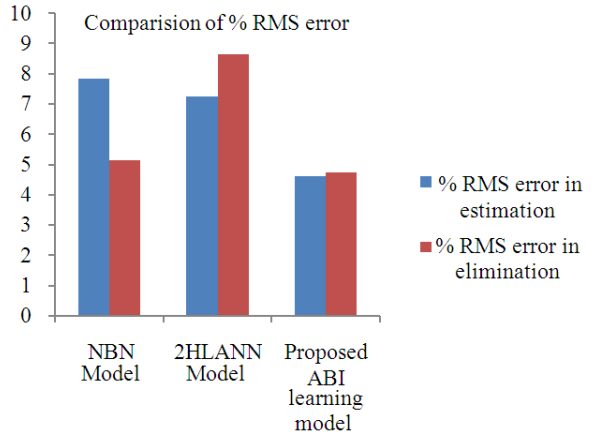

Fig. 2. RMS Error in proposed and existing models

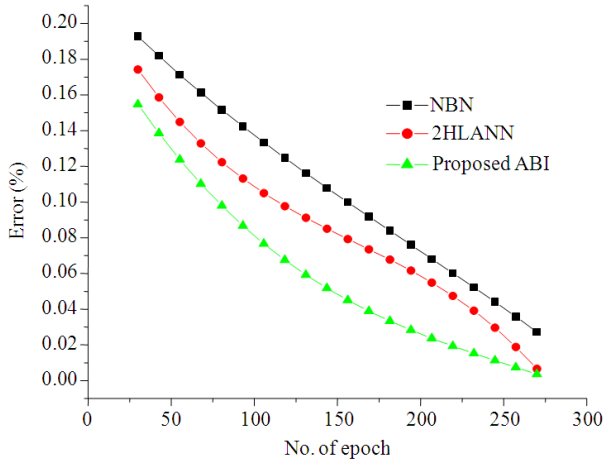

Fig. 3. % Error in proposed and existing models

The estimation is improved around 25% than NBN and 33% than 2HLANN. Similarly the elimination is improved around 30% than NBN and 33% than 2HLANN. The accuracy of the proposed system also improved which is shown in the error rate and learning rate based on the epoch.

Figure 2 shows the graphical representation of the performance of proposed and existing models. **Figure 3** and **Table 1** shows that the proposed learning model reaches the performance 0.010 in 250 epochs (number of iterations), whereas the existing NBN Model reaches only 0.04 and 2HLANN reaches only 0.025 respectively, which is lesser than the proposed system. Therefore, the proposed ABI learning is more optimal than existing models.

4. CONCLUSION

The objective of our work is to efficiently mine knowledge from the unstructured text documents. In order to mine textual documents, text mining is applied. The text mining is the sub-domain in data mining. The performance of VMM is improved using ABI learning algorithm. The proposed ABI learns the synonymy better than the existing systems. The proposed ABI learning method improved estimation, elimination and accuracy of the system. The estimation is improved around 25% than NBN and 33% than 2HLANN. Similarly the elimination is improved around 30% than NBN and 33% than 2HLANN. The accuracy of the system also improved the error rate and learning rate based on the epoch. From the result, it is concluded that the proposed VMM with ABI learning algorithm is proved better result than existing and notable recent works in document clustering field of domain.

5. REFERENCES

Agrafiotis, D.K. and H. Xu, 2002. A self-organizing principle for learning nonlinear manifolds. Proc. Nat. Acad. Sci. USA, 99: 15869-15872. DOI: 10.1073/pnas.242424399

Cai, D., X. He and J. Han, 2005. Document clustering using locality preserving indexing. IEEE Trans. Knowl. Data Eng.., 17: 1624-1637. DOI: 10.1109/TKDE.2005.198

Cai, D., X. He and J. Han, 2011. Locally consistent concept factorization for document clustering. IEEE Trans. Knowl. Data Eng.., 23: 902-913. DOI: 10.1109/TKDE.2010.165

Cheng, D., R. Kannan, S. Vempala and G. Wang, 2006. A divide-and-merge methodology for clustering. ACM Trans. Database Syst., 31: 1499-1525. DOI: 10.1145/1189769.1189779

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A Review. ACM Comput. Surveys, 31: 264-323. DOI: 10.1145/331499.331504

Koteeswaran, S., P. Visu and J. Janet, 2012a. A review on clustering and outlier analysis techniques in datamining. Am. J. Applied Sci., 9: 254-258. DOI: 10.3844/ajassp.2012.254.258

Koteeswaran, S., J. Janet and E. Kannan, 2012b. Significant term list based metadata conceptual mining model for effective text clustering. J. Comput. Sci., 8: 1660-1666. DOI: 10.3844/jcssp.2012.1660.1666

Koteeswaran, S. and E. Kannan, 2013. Analysis of Bilateral Intelligence (ABI) for textual pattern learning. Inform. Technol. J., 12: 867-870. DOI: 10.3923/itj.2013.867.870

Kotsiantis, S. and P. Pintelas, 2004. Recent advances in clustering: A brief survey. WSEAS Trans. Inform. Sci. Applic., 1: 73-81.

Lebanon. G., 2006. Metric learning for text documents. IEEE Trans. Patt. Anal. Mach. Intell., 28: 497-507. DOI: 10.1109/TPAMI.2006.77

Mkadem, F. and S. Boumaiza, 2011. Physically Inspired neural network model for rf power amplifier behavioral modeling and digital predistortion. IEEE Trans. Microwave Theory Techniques, 59: 913-923. DOI: 10.1109/TMTT.2010.2098041

Ng, R.T. and J. Han, 1994. Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th International Conference on Very Large Data Bases, (LDB' 94), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 144-155..

Zhong, S. and J. Ghosh, 2003. Model-based clustering with soft balancing. Proceedings of the 3rd SIAM International Conference on Data Mining, Nov. 19-22, IEEE Xplore Press, pp: 71-82. DOI: 10.1109/ICDM.2003.1250953

Wilamowski, B.M., 2009. Neural network architectures and learning algorithms. IEEE Indust. Electron. Magaz., 3: 56-63. DOI: 10.1109/MIE.2009.934790

Zha, H., C. Ding, M. Gu, X. He and H. Simon, 2001. Spectral relaxation for k-means. Neural Inform. Process. Syst., 14: 1057-1064

Zhu, X., Z. Ghahramani and J. Lafferty, 2003. Semi-supervised learning using gaussian fields and harmonic functions. Proceedings of the 20th International Conference on Machine Learning Aug. 21-24, Washington, DC USA., pp: 912-919.