ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse107952020.pdf
https://doi.org/10.30534/ijatcse/2020/107952020

Multinomial Classification of "Hete-Neurons" in Heterogeneous Information Networks

Sadhana Kodali [1], Madhavi Dabbiru [2], B Thirumala Rao [3]
[1,]PhD Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[2]Professor, Department of Computer Science Engineering, Dr.L.Bullayya College of Engineering, Visakhapatnam, Andhra Pradesh, India.
[3,] Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

ABSTRACT

The real-world objects communicate and exchange data and can have meaningful relationships between them which lead to the field of research called Heterogeneous Information Networks. These object behaviour and attributes are modelled as "Hete-Neurons" which is proposed in this paper. To model this object interconnectivity and Multinomial Classification the cognitive level of mapping of the queries in the examination by using Bloom's Taxonomy is considered and interesting results are identified. The comparative study held on previous models which used the Bloom's Taxonomy showed an accuracy of 89 % but the proposed approach presented a higher accuracy and also gave a new insight into the perception of a neuron as "Hete-Neuron" which is heterogeneous. A multi-class Classification approach is experimented with a dataset which contains 1050 queries from well reputed university exam papers and showed a higher accuracy rate with 99.86 when tested with the sigmoid function. The same dataset is also experimented using the well-known softmax function which gave an accuracy of 100%.

Key words: Hete-Neuron, Multinomial Classification, Bloom's Taxonomy, Heterogeneous Information Networks.

1.INTRODUCTION

In Machine Learning the classification task can be of two types:1.Binary Classification 2.Multiclass Classification. Binary Classification is the act of classifying objects to one of the two given classes. While Multiclass Classification classifies the objects to more than three classes. The Multiclass classification is also called Multinomial Classification. These two terms will be interchangeably used throughout this paper. The act of predicting the class label of the given objects from a semantically connected network of objects based on certain metrics is a challenging issue. The main aim of this paper is to identify the interconnecting objects by text analysis and classifying the objects using Multiclass or Multinomial Classification. The objects may have diversified relationships and the objects which are interconnected may not be the same and are heterogeneous in nature.This interconnectivity of different objects with semantically and structurally diversified relationships is called a Heterogeneous Information Network(HIN)[1].The objects which are connected may also be homogeneous in nature.This paper focuses on observing the paths which are known as meta-paths that lead to the meaningful relationships between these objects, predicting the class label using machine learning algorithm and to map these objects to Multiclass. According to revised Bloom's Taxonomy cognitive levels are classified into six levels namely 1.Remember 2.Understand 3.Apply 4.Analyse 5.Evaluate 6.Create.Therefore, the inputs are finally mapped to more than one class in the above six classes.

To classify the heterogeneous objects to multiclass and for experimentation purposes the questions in the examinations according to Bloom's taxonomy play a major role which estimate the quality of the paper. A few research papers came into existence to do the classification of questions in Bloom's taxonomy but these were the best classification models to classify the questions for single class classification. The proposed method classifies the method using multinomial classification and also the test results showed higher accuracy when compared to the previous methods. When a question is framed in the University exams the verbs used may be multiple. If a question consists multiple verbs and classifying it to one single class may miss the other verbs which are neglected without classifying. This paper also presents an approach of understanding the concept of heterogeneous networks using text classification. A heterogeneous information network is the interaction between dissimilar objects and the objects are interconnected by distinct relationships between them. In this paper the authors analyse the different subjects of the universities belong to one category of object. In each subject the examination is conducted and there may be different kinds of examinations like the semester examinations , regular and weekly assessments therefore an examination is a kind of object. The queries in the exams derive the quality of the

student which leads to an outcome. The verbs in the queries are treated as one category of objects. Mapping these verbs to the cognitive domain of the bloom's taxonomy is very helpful in the assessment methodology. As shown in Figure-1 the university may have various programs example different streams and in turn each program has a set of courses which are the different subjects taught to the students per semester. The assessment for the subject is made by conducting examination and there may be different kinds of examinations like day-to-day evaluations, weekly evaluations, monthly evaluation and end-semester examination. The examination contains queries based on the marks and the queries are categorised as question 1,2, so on. The verbs in the queries are mapped to the Bloom's taxonomy. Based on the key words used in the query it can be classified in the Bloom's Taxonomy.

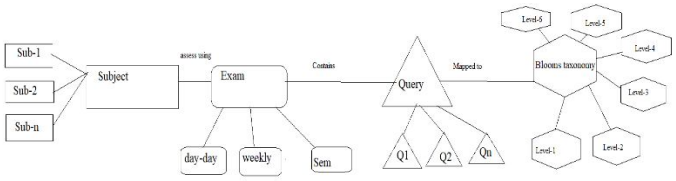

Figure 1:Heterogeneous Object interaction using Query Mapping to Bloom's Taxonomy.

The Figure1 indicates how heterogeneous components interact with each other and the different semantic relationships between them.

The paper is organised by discussing the literature survey in the second section, the proposed methodology in section 3,experimentation results in section 4 followed by conclusion and future scope in section 5.

2. LITERATURE SURVEY

A neural network uses a set of algorithms to find out the underlying relationships in the data just as the human brain does. The Heterogeneous Information Network also has underlying semantic relationships between data objects. To identify the semantic relationships in graphs which consists of heterogeneous objects, a Graph Neural Network [2][3] was proposed that explained the framework for neural convolution in the graphs which led to graph learning but the drawback of this method is its low efficiency. In a research work [4], non-local dependencies are embedded and the approach is named as Non-local attention learning, which uses a hierarchical mechanism to predict the long range dependencies and minimises the localization problem of identifying the node wise objects versus another research focussed on semantic objects in a hierarchical heterogeneous information network which is named as Heterogeneous graph attention Networks(HAN) [5]. A learning method called relational triplet network approach to identify the detailed semantic embeddings of heterogeneous information networks is proposed in one more research[6], which found to be very useful for link prediction and classification by traversing the meta-paths. By pre-training and fine tuning of heterogeneous information network to represent it in a low dimension space a method called PF-HIN is used in [7] where a node sequence pair is used for link prediction but uses only single node sequence for the purpose of classification.

The early research of heterogeneous neural network was done by using the artificial neural network model which learns from the examples and is modelled to tolerate the partial failures [8] .The use of artificial neural networks for analysing meta-paths from the heterogeneous information network found to show results with high accuracy.

The authors of this paper choose to apply multi-class text classification for classification of question papers from the universities which require cognitive level of assessment using the Bloom's Taxonomy which caught the interest of many schools and university for the assessment mechanism. Few interesting papers to classify the questions using Bloom's Taxonomy gained acceleration these days but the work did not identify the multi-class classification in the question papers and the concept of heterogeneous object interaction was not focused in these papers[9,10].In a research paper[9] Machine Learning techniques are used to classify the questions using Logistic Regression and Linear Discriminant analysis with an accuracy of 83.4%.A fuzzy classification technique was developed in [10] and GUI based application called bloom analyser was created to classify the verbs in the question paper but it was single class classification though it has very enhanced features like creating e-libraries for QMS,the test results conducted by the experiments resulted in a confidence of 96.2% and the accuracy calculated was called as percent correct and showed 88 % for bagging,85.2 % for Bayes Net and 88.2-88.3 % for Random forest and Random Tree respectively. In [11] the knowledge based semantic measurement is also studied to understand Information retrieval and Natural Language Processing. The authors coined the term 'Hete-Neurons' which serve as inputs to train the neural network. 'Hete-Neurons' indicate the heterogeneous object interaction which is justified in the section 3. This proposed approach has greater accuracy when compared to previous approaches

3. METHODOLOGY

The artificial neuron which receives the input from various sources is perceived as homogeneous but which actually is not. The reason for this is all the real world inputs are obtained from various sources have a mixture of variables like numeric and qualitative. The authors in this paper presented a term "Hete-Neuron" which represents that a neuron can be more than one type may it be categorical, ordinal,nominal. The perception of neuron as a heterogeneous object can lead into traversing the various meta-paths of a heterogeneous information network modelling it as an artificial neural network in a framework

which contains heterogeneous inputs and weights. The proposed method works with a three layer Artificial Neural Network which has an input layer, hidden layer and an output layer as indicated in Figure 2.The data pre-processing is done using the sigmoid function to normalise the values. The queries which are the inputs are later transformed to bag-of-words.The sigmoid function is given in Equation 1 as :

$$Sigmoid(x) = \frac{1}{1+e^{-\beta x}} \qquad (1)$$

Where x is the word in lowercase from bag_of _words and β is a constant used to normalise the given inputs in (1). To normalise the values error-rate was iterated and finally fixed at 0.2.The choice of sigmoid function is made due to the reason that the obtained class labels are not mutually exclusive and can be chosen from one of the obtained output Class Labels. The sigmoid function is more specific for binary classification. But, the aim of this paper is to do multiclass or multinomial classification the better option would be the softmax function instead of the sigmoid. The softmax function calculates the probability of the target Class for each possible values.

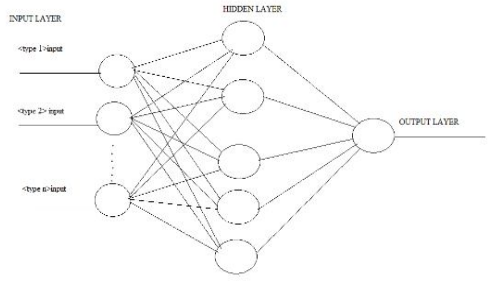

Figure 2: A three layer neural network

The softmax function is given in Equation 2 as :

$$Softmax(x) = \frac{e^{Vi}}{\sum_j^o e^{pj}} \qquad (2)$$

The Equation 2 is used to calculate the softmax function, where the V_i in the numerator is the input vector and the nature of the inputs are real values. The p_j in the denominator gives the output probability distribution of 'o' probability that is proportional to the exponential of the input. The dataset is experimented on both the sigmoid function and the softmax function in (2).The experiment results obtained for softmax function showed 100% accuracy which indicates the use of softmax function for multi-class classification of a Hete-Neuron shows optimal result.

The total number of training sentences is 140 sentences in the training data to classify the verbs according to Bloom's taxonomy. The method to perform the classification is defined. The input dataset contains the queries in sentences format. Later these sentences are split to bag_of_words which is an array of 0's and 1's.The presence or absence of the keywords defined in the Bloom's Taxonomy the weight of every input sentence is computed and normalised with sigmoid function. The procedure can be clearly explained from Figure 3.

Algorithm Classify_Hete_neuron

Step 1: Input Sentence

Step 2:Convert the sentences into bag-of-words

Step 3: Compute the synaptic matrix and assign the weights.

For j in iterations

Assign the trained value to input layer

Compute the sigmoid function to hidden layer to using input layer synapse.

Normalise the error rate;

Update Synapse_weight for input layer;

Update Synapse_weight for hidden layer;

return result;

Step 4:Fix the error_threshold

Step 5:for r in result check if r>error_threshold

If true sort(result)

For every Class obtained:

Print(ClassLabel)

Figure 3: Algorithm Classify_Hete_neuron

The approach of Hete-Neuron can be clearly understood from the Figure 4 which has a heterogeneous inputs at the layer 1 and the hidden layer which consists of adjustable size of the number neurons such that the values are normalised and the synaptics are obtained. The third layer consists of the output layer which has the final multi-class Class Labels.Traversing from each of the input in the Hete-Neuron gives the meta-path to the output.

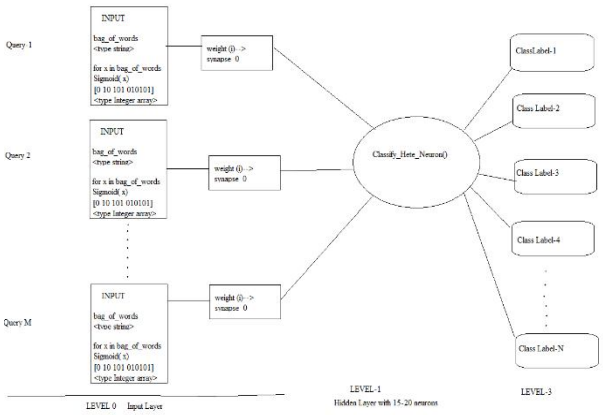

Figure 4: A Hete-Neuron adapting a three layer ANN model.

4. EXPERIMENTATION

The experimentation is executed in a Cloud using Google Colab Notebook which is very helpful to conduct experiments related to Machine Learning.The necessary

packages like the NumPy and related Tensor Flow and nltk packages are imported.The input file which contains queries from a well reputed university which contains 1050 questions is given as an input in the form of csv file. The fields in the input file contains the query number ,the Subject to which the query belongs, the query which is a string and is later tokenized to bag_of_words during pre-processing.The dataset contains the following fields using given in Table 1.

Table 1: Dataset description

No of Rows	1050
Serial Number	Integer
Subject Name	String The name of different courses handled by the college or universities
Query	heterogeneous data like text,numbers,special symbols,equations.

Later the training data with 140 sentences is trained to define the six cognitive levels of the Bloom'sTaxonomy. The Figure 5 explains how each sentence is classifies according to the Bloom's Taxonomy.

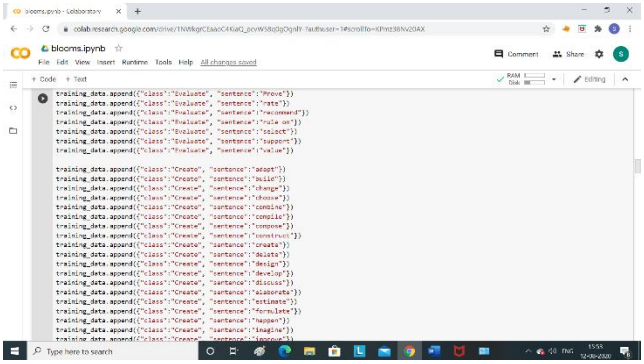

Figure 5:A sample of training key words.

The processing time and the number of neurons in the hidden layer can be clearly understood from Figure 6.

```
Training with 20 neurons, alpha:0.1, dropout:False
Input matrix: 140x120    Output matrix: 1x7
delta after 10000 iterations:0.05038304198159425
delta after 20000 iterations:0.04950161514707104
delta after 30000 iterations:0.04912773618298845
delta after 40000 iterations:0.04890982456967353
delta after 50000 iterations:0.04876331605119878
delta after 60000 iterations:0.048656352647046874
delta after 70000 iterations:0.04857393780293187
delta after 80000 iterations:0.048507975135444435
delta after 90000 iterations:0.048453663692983956
delta after 100000 iterations:0.04840795463877477
saved synapses to: synapses.json
processing time: 45.319045543670654 seconds
```

Figure 6: The number of neurons involved in the hidden layer and the processing time.

The test query can be given by using the method classify(query).After the query is classified it gives the accuracy along with the Class Labels. The Figure 7 clearly demonstrates the Class Labels and the accuracy obtained for each query. The accuracy computed is around 99.986% which has a better accuracy when compared with the other models studied in [8] and [9].

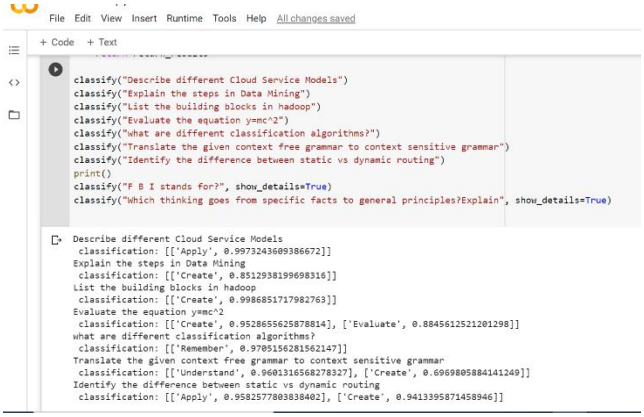

Figure7: Classifying the Test queries and their corresponding accuracy measure.

With the given dataset as an input the logistic regression model used in [8] is executed which gives an accuracy of 89%.Compared to the logistic regression model the ANN model on heterogeneous information networks has a better accuracy as shown in Figure 8.

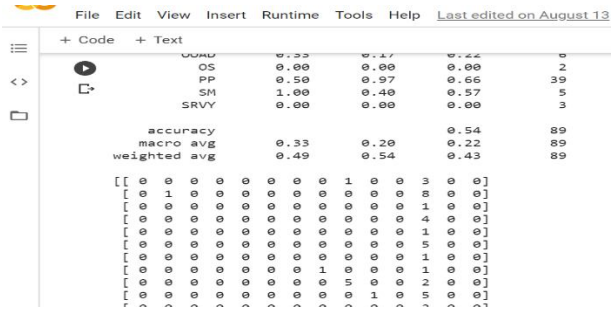

Figure 8: The Logistic Regression applied on the dataset with 1050 queries which gave an accuracy of 89 %.

Similarly, the Figure 9 below shows how softmax worked on the same dataset to give an accuracy of 100% with an error rate of 0.2.

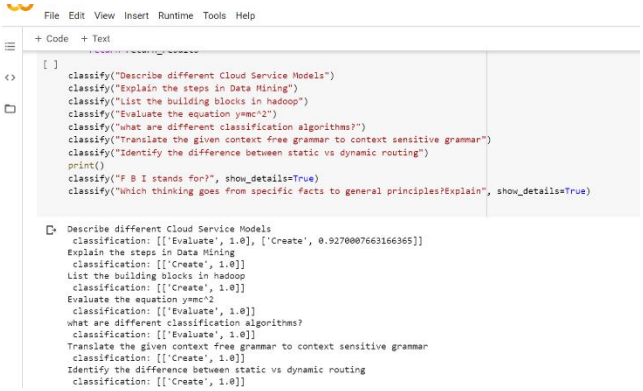

Figure 9: Multinomial Classification of Hete-Neurons using softmax.

Table 2 demonstrates the accuracy of the proposed Hete-Neuron multinomial classifier when compared to logistic Regression approach. Also represents the resulted accuracy with the usage of sigmoid function versus softmax function.

Table 2: The comparative calculations with LR and Hete-Neuron based multi class classification.

Name	Logistic Regression	Multi-Class classification in HIN using sigmoid	Multi-Class classification in HIN using softmax
accuracy	89%	99.86%	100%

5.CONCLUSION

Every object interaction can be modelled as a Heterogeneous Information Network.The HIN can be better analysed by mapping it to a neural network approach. Traversing the multi-paths and identifying the neurons can produce interesting results and these neurons can be modelled as 'Hete-Neurons'. There may sometimes be a chance where the object can be mapped to more than one class this multiclass Classification can be a useful approach and can have many applications in analysing the social networks, biological networks and many more. Observing the semantic and structural relations between the objects will be very important task to trace out the 'Hete-Neurons'. The identification of Hete-Neurons can promote new avenue of research.

REFERENCES

1. Chuan Shi,Yitong Li,Jiawei Zhang,Yizhou Sun,PhilipSYu,**A survey of heterogeneous information network analysis** IEEE transactions on Knowledge and Data Engineering",Volume 29,Issue 1,Jan 1 2017.
2. W. Hamilton, Z. Ying, and J. Leskovec **Inductive representation learning on large graphs** in Advances in Neural Information Processing Systems, 2017, pp. 1024–1034.
3. Noussaima EL Khattabi, Fouad yakoubi, Hajar Sahbani , Mohamed El Marraki,**Computing the number of spanning trees in the Wheel multiple graph,** IJATCSE,Vol8,No 4 ,August 2019.
4. Yuxin Xiao, Zecheng Zhang, Carl Yang, and Chengxiang Zhai **Non-local Attention Learning on Large Heterogeneous Information Networks,**2019 IEEE International Conference on Big Data (Big Data), *Los Angeles, CA, USA, 2019, pp. 978-987*
5. X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu **Heterogeneous graph attention network** in Proceedings of the 2019 World Wide Web Conference (WWW), 2019.
6. Xiyue Gao,Jun Chen, Zexing Zhan,Shuai Yang,**Learning heterogeneous information network embeddings via relational triplet network,**Neurocomputing,2020.
7. Yang Fang, Xiang Zhao, Weidong Xiao,**Exploring Heterogeneous Information Networks via Pre-Training,**IEEE,July 2020.
8. Lluís Belanche, **Heterogeneous neural networks: theory and applications,** March 2009.
9. Jain M., Beniwal R., Ghosh A., Grover T., Tyagi U.**Classifying Question Papers with Bloom's Taxonomy Using Machine Learning Techniques** Advances in Computing and Data Sciences. ICACDS 2019. Communications in Computer and Information Science, vol 1046. Springer, Singapore.
10. Fouad Jameel Ibrahim AlAzzawi et al., **Fuzzy Analysis Model for Classifying Exams Questions in Learning Quality Management System Based on Bloom's Taxonomy Verbs**Associative Ontology Systems ,October 2019.
11. Ali Muttaleb Hasan, Noorhuzaimi Mohd Noor , Taha. H. Rassem, Ahmed Muttaleb Hasan **Knowledge-Based Semantic Relatedness measure using Semantic features ,** IJATCSE ,Vol 9 No 2 April 2020