Rank Prediction for Articles and Conference Papers using Machine Learning Techniques.

P.Subhashini, K. SatyaVijaya, G.S.K.Bhargav, V.Vjiay Kumar

Abstract: Searching for an optimal article which was given highest and best priority is quite harder based on requirements. Ranking is one of the best measure or a method to get the best rated and optimal article or a conference or a research paper through this huge Internet World. As Technology been increasing day by day Artificial Intelligence is the first step to get through any problem for a solution Machine learning is also an important aspect of Artificial Intelligence. Machine Learning is best known for classifying, categorizing and predicting. Rank prediction can be done through many different algorithm implementations in machine learning. But choosing the best is important for accurate results. This paper gives the most accurate results of algorithms that can be used for rank predictions for articles. To simplify and resolve this problem, solutions were given in many different ways but to achieve accuracy is necessary, in previous models this is given using supervised learning only. We proposed this research work with perfect results using both supervised and unsupervised learning. Neural Networks is the best algorithm in supervised learning for classifying and predicting within data. In unsupervised learning we used K-means clustering because of grouping the data. This work helps the user(s) for optimal search of an article and also gives a competitive spirit for author to get into the top, totally this is implemented using Machine Learning Techniques of Neural Networks, K-Means Algorithm which is a mixture of supervised and unsupervised learning for predicting ranks.

*Keywords: Rank Predictions, Neural Networks,
K-Means, Top Ranked Articles, Accuracy.*

I. INTRODUCTION

Because of vast improvement of information and technology advances as of late, ranking analysis of an article from expanded fields of research has become an extremely indispensable errand because of the exponential development in every day production in lives and meetings, outperforming fifty million papers. Additionally, foreseeing the since quite a while ago of any system identify another task that we come across regularly, Anyway each issue is a case that will be addressed during the investigation or analysis by giving the new rank for a paper.

Manuscript published on November 30, 2019.
* Correspondence Author

P. Subhashini*, Department of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India.
K. SatyaVijaya, Department of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India.
G.S.K Bhargav, Department of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India.
V. Vijay Kumar, Department of CSE, professor at Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India.

Utilizing AI in analysis of positioning or ranking of papers could be a significant research heading, because of it contains unmistakable classifications of supervised learning algorithms with pertinent in expectation. Between the most imperative calculations utilized in classifiers, we choose to figure out with multilayer perceptron program of algorithms, Neural Network Classifiers and k-means clustering. These methods speak of cell coordinate (network) with the hubs and edges as paper or articles and references (citations) between them.The principle explanation behind this determination is that system analysis of paper network could be a case of connection wherever the researchers teams up with various network communicators in technological domains in order to understand their objectives.The associated work gives a vision on the methodologies and ways for the arrangement of research project papers, and that is grouped into 2 significant coordinates, the essential coordinate is ranking it with a reference predict and second is about the rank in technical connection. The Prediction of Ranks for a set of Research or Article papers is an automatic representation of the longer term for anevent of anyframework, particularly in searching. During this project, we tend to portray one in all programmed arrangement procedures applied, to look into a supervised learning task [1]. All through the strategy, we tend to build up the most choices that square measure utilized as keys to assume a major job as far as anticipating the prediction of new rank underneath the learning arrangements. In this paper,our mission is to predict the rank for an article or conference paper and check the highest one and find accuracy with the algorithms used and Check whether the accuracy rate is high as required and as proved previously. This paper is further divided as following, Section-II is about the related work that describes the gather and research information of the topic. Section-III is about the methodology used to solve and rank the data. Section-IV represents the implementation and processing of data and algorithms. Section-V deals with the Results and Experiment. Section-VI is all that says the paper review and conclusion on our research. Section-VII is of References and representations.

II. RELATEDWORK

The Ranking for any scientific research article or research paper,journals, Conference papers has been a recent interesting topic on the net [6]. In recent times, there been some other projects which relate to this. So, as a research and contribution from us we used supervised learning Techniques in Machine learning.

Supervised Learning:

It is defined that the machine learning undertaking of learning a work that maps its partner or neighbor input to relate outcome that upheld model information with in-out sets [4].It induces a work from labeled training information comprising of an assortment of instructing models. In supervised learning, every model might be a takeat comprising of next input object (regularly a vector) and an optimal output [4].Supervised learning algorithmic rule resembles instructing information and produces an inductivework, which might be utilized for mapping new example associate tooptimum state of affairs that can yield the algorithmic rule to properly verify the category labels for unseen instances.

The main algorithmsrelated are as follows:

- Neural Networks.
- Linear Regression
- Naïve Bayes
- SVM- Support Vector Machine etc.

Neural Networks:

It is a concept in artificial intelligence where this algorithm is a computational learning framework that uses a system of capacities to comprehend and interpret an information contribution of one structure into an ideal outcome, for the most part in another structure [11]. The idea of this system was motivated by human science and path neurons of the human mind work together to comprehend from human detections. The Neural capacities are connected with this information, covered up and yield layers [1].

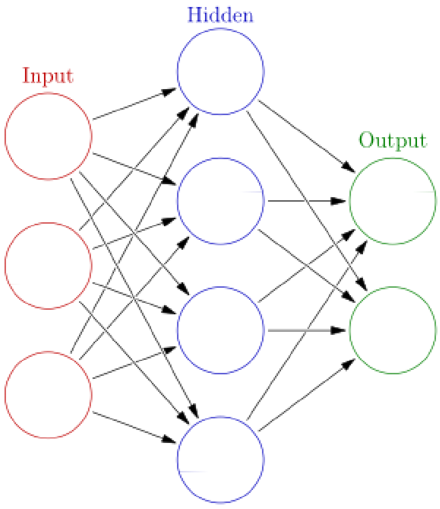


Fig.1. Example of Neural Network

Unsupervised learning:

Unsupervised learning is the preparation of machine utilizing data that is neither arranged nor marked and enabling the calculation to follow up on that data without direction. Here the assignment of machine is aggregate unsorted data as per similitudes, examples and contrasts with no earlier preparing of information [1].

• K-means Clustering

K-means Clustering:

K-means Clustering plans to divide n objects into k groups inside which each object has a place with the group with the nearest mean [13]. This method delivers explicitly k entirely unexpected group of most prominent data of qualification.

- **K-means++:**K-Means++ begins with the assignment of one cluster focus to other cluster focus from the given clusters data.

In Fig.2 we can see the centers taken as yellow circles in each cluster for initializing.

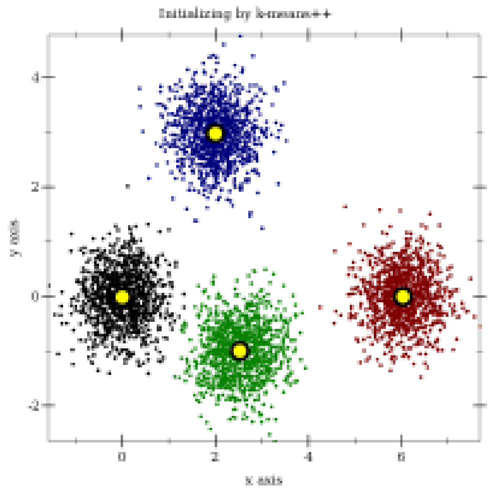


Fig.2. Example of K-Means++

As supervised learning is the best approach for not having any mislabeled data. As each and every article has their own attributes accordingly there are no mislabeled data in the dataset. It is also most effective and accurate compared to unsupervised and Semi-supervised learning [3]. So, we have chosen both supervised and unsupervised learning where both combined can give easy and accurate results.Compared to rest other supervised learning such as Hidden Markov model, SVM (Support Vector Machine) [4], Linear Regression, Naïve Bayes and others neural network is more suitable and better algorithm for huge datasets. Same as in unsupervised learning helps to solve by division of data through Clustering or Classifications. Hence we choose to take K-means Clustering.In Previous Model it was proposed that algorithms like K-means and Fuzzy c-means algorithms are used together for ranking of whole dataset. It results the same accuracy, but complex we can use K-Means++ which helps to initialize first and divide into clusters for a better outcome.However, there are other previous models used for ranking such as Vector Space Model, Boolean Ranking Model, Okapi BM25, Page rank etc. Next Section explains about the methodology used for ranking.

III. METHODOLOGY

The methods we used in this paper are different while compared to others. We used tokenization and splitting to check the different words and popularity of the paper.

From collection of all the data as this paper is based on clustering and networking each paper hub x joins with next paper hub x as an attribute value in them and also with other node y as popularity value with each paper then the count of terms were taken and weighted with number of duplicates. Tokenization is used to convert each article or paper astoken and split them to find the duplication. After then these were formed into clusters.

These methods were generally taken from Natural Language Training Kit – NLTK. We performed this experiment using Python language. Python is a very reliable and experimental language with Machine learning.

There are different Libraries used for implements the packages and function to result successfully.

- The Article or paper with high citation value will be given the highest rank.
- Also depends on duplication and no split value.
- The popularity and review also matters.

Python is a translated raised level programming language.

As interpreted it is a direct execution and suitable for exploring different packages. After the collection of data about papers the packages explore and perform the implemented algorithms and steps.

IV. IMPLEMENTATION

The basic steps for this experiment is given in the below Block diagram:

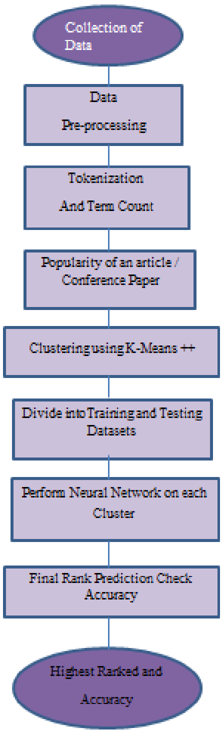

Collection of Data
Data Pre-processing
Tokenization And Term Count
Popularity of an article / Conference Paper
Clustering using K-Means ++
Divide into Training and Testing Datasets
Perform Neural Network on each Cluster
Final Rank Prediction Check Accuracy
Highest Ranked and Accuracy

Fig.3. Block Diagram

On the first node we can see,

1. Collection of Data:

We have collected the datasets from sources like UCI repository Datasets. The Dataset for this project may vary every time based on their attributes. So it's not constant hence, it's a myth for this paper.

As of the collected data we have about 210 tuples (articles) and 7 attributes (columns).

Each attribute defines the paper by their citation and value of the paper

2. Tokenization and Term Count:

We used Tokenization to resolve the text into words in python. The main attribute was split into tokens. With split tokens we take the term count and weight it to check the popularity.

Tokenization is an easy method in python to implement NLTK libraries.

3. Popularity of an article / paper:

The popularity of a paper depends on its citation, review, authors, importance of the article. Popularity of the paper can be determined using the term count and duplication of words and pages.

The popularity is the most important attribute that defines the Rank of a paper.

4. Clustering :

Clustering is defined as grouping of the data which are similar to each other. Clustering can be implemented by different methods. From chosen we used K-Means clustering for dividing the data into groups. We have totally of 6 clusters formed on the basis of similarity.

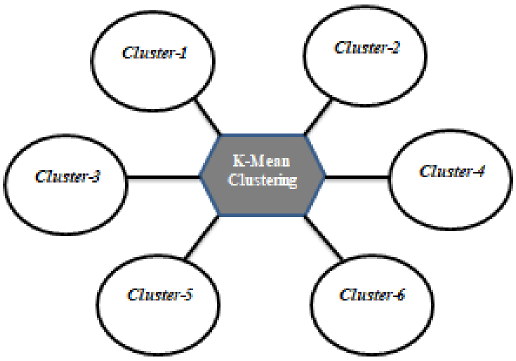


Fig.4.Clustering

Each Cluster contains the similar set of papers in it.

But this may vary with attributes and article details. So defining each paper with an initial rank using K-Mean++. So that categorizing the data about the articles these clusters were further taken to Neural Networks algorithm and then defined the next procedure.

5. Dividing into Testing and Training Datasets:

The clusters taken into Neural Networks were divided into training and testing datasets.

The Neural Networks uses different layers for sorting the data and divide into training and testing datasets.

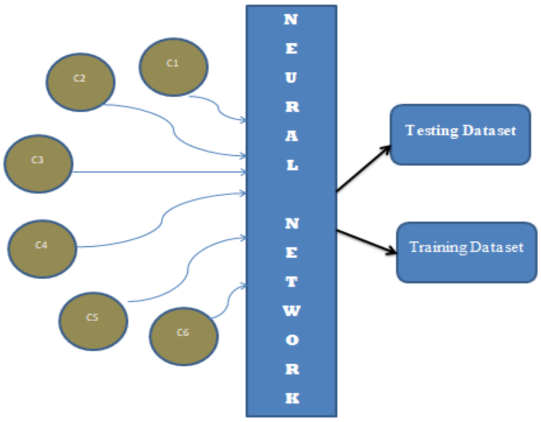


Fig.5

Here in Fig.5 C1, C2, C3, C4, C5, C6 are the clusters. These training and testing datasets are later taken into neural network algorithm and further techniques are applied.

6. Neural Networks Performance on datasets

The process of neural networks is again performed using training and testing datasets.

Built Neural Networks for training dataset first and predict with testing dataset. The data in those datasets were carried out by comparing one article with the rest of the articles with relevant formula and algorithmic rules.

- With the input to the hidden layer each article is mapped to other and checks the value of their citation and popularity and gives the best out each and results accordingly. In such a way, that each article is performed and ranked respectively.

7. Rank Prediction and Accuracy:

After all the algorithms application the rank is predicted as through the above steps. The ranks are not unique for every paper because their might be same features for different papers. Computing training and testing datasets we get the predictable ranks and accuracy.

8. **Plot the Graph** accordingly and the highest point owns the highest rank.By 4 folds of the dataset we can find the accuracy of the experiment.Next section is about the results of the experiment.

V. RESULTS AND ANALYSIS

Overall the performance of the experiment is discussed here. Final output we represent is as follows;

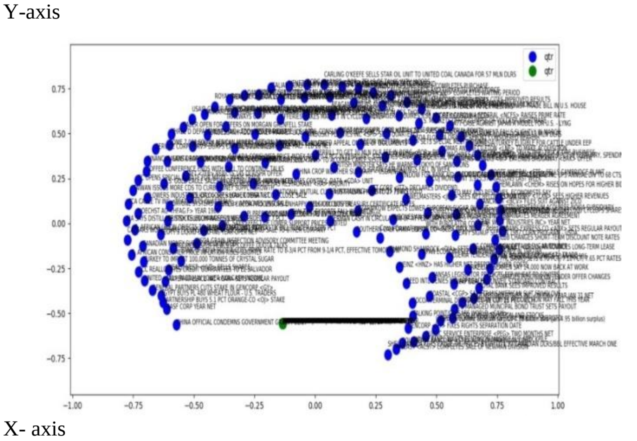


Fig.6

As this experiment is done on large dataset, there are more than one article ranked same. This experiment gives the outcome as rank prediction for a set of articles or papers and predicts the highest rank.

In Fig.6 graph the X and Y axis determines the citations and popularity ranking. The peak point in the graph determines the highest rank among all the papers.

The accuracy calculated is given as;

FOLD (out of 4)	Accuracy train %	Accuracy valid %	n train	n valid
4-Jan	84	90	158	52
4-Feb	86	92	158	52
4-Mar	86	76	158	52
4-Apr	86	76	158	52

Fig.7

The Above table defines the accuracy rate for each fold of cross validation of rank prediction.

FOLD (out of 4)	Accuracy train %	Accuracy valid %	n train	n valid
4-Jan	84	90	158	52
4-Feb	86	92	158	52
4-Mar	86	76	158	52
4-Apr	86	76	158	52

Fig.8

In the above table Fig.8 it can be observed that the accuracy varies in both training and testing validation. In the first fold the training accuracy was low compared to other folds and testing validation of accuracy in second fold was high with other folds.Hence this experiment analysis says that each fold determines different percentages or same but the overall average is to be given optimal as the research says.

In research of this paper we can say that the average accuracy should be given as follows;

1. Accuracy <= 0% : Very low or No accurate outcome
2. 40% <= Accuracy <= 80% : At medium level
3. Accuracy > 80%: High in accuracy level.

Therefore, the average accuracy that is calculated through this experiment is

Average of Acc_train	Average of Acc_valid
85.50 %	**83.50%**

Fig.9

This accuracy calculated in the experiment is high in level. Hence rank prediction is done with high accuracy.

VI. CONCLUSION

In this paper we proposed a mixed usage of supervised and unsupervised learning algorithms of k-means, k-means++ clustering and neural network classifier to predict the ranking of the scientific research papers. unlike the other methods, usage of k-means++ algorithm gives us the evenly and correctly divided clusters which in the end helps to predict the new rank using neural networks. This method is more reliable and concrete as the methods used are not based on luck but accuracy. We process the data by considering the key attributes about the paper or article like citations, downloads, authors etc., as these cannot be overtaken by any other attributes. We also used some NLTK (Natural Language Training Kit) methods to saturate the data inorder to make it easier in further computations. Finally we got the accuracy as 85.5 % which is a good result. Hence we can as that the prediction accuracy doesn't change even if we use both supervised and unsupervised learning.

We plan to test this method on different datasets as the different types of datasets gives different outcome. This is to make sure that it gives us the same accuracy levels when also used with supervised learning and unsupervised learning. As of now, we are pretty sure about the robustness of the method we used, as the algorithms are pretty clear and balanced.

REFERENCES

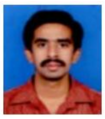
1. Multilayer perception: architecture optimization and training Int. J. Interact. Multimed. Artif. Intell.4 – 2016 – .Ramchoun, M. Amine, J. Idrissi, Y. Ghanon, M. Ettaouil.
2. A new extrapolation method of Page Rank computations by X.Tan, J. Comput. Appl. Math., 313(2017), pp. 383-392.
3. Supervised learning and co-training, Algorithmic learn. By M.Darnstadt, H.U. Simon, B. Szorenyi.- 2013
4. Hassan SAyyadi, LiseGetoor – Futur Rank: Ranking Scientific Articles by predicting their future page rank.
5. Predicting Win percentage and winning features of NBA teams CS229 final report evangiatra, nattapoomasavarcongchai.
6. M. Pradeepini, Anila G.2017, “Study of perdiction algorithms for selecting appropriate classifier in machine learning”, Journal of advanced Research in Dynamical and control systems, vol. 9,no. Special issue 18, pp.257-268.
7. “Least square regression for prediction problems in machine learning using R” –2018, by Anial, M.Pradeepini, International journal of engineering and Technology (UAE), vol. 7 no. 3.12.
8. “Assessment and analysis of software reliability using machine learning Techniques”, by Krishna Mohan, N. Lavanya, G. Yoshitha, M.L.n. Krishna priya . A – 2018, Internation journal of Engineering and Technology (UAE), vol.7 no. 2.32 Special Issue 32, pp.201-205.
9. “Data Classification and personal care management system by machine learning approach” by G. Pradeepini, G. Pradeepa, Tejanagasri, B. Gorrepati, S.H. 2018, Interntional Journal of Engineering and Technology(UAE) , vol. 7 no. 2.32 Special Issue 32,pp.219-223.
10. Chandana . K, Prasanth . Y, Prabhu Das . J,2016- “A decision support system for predicting diabetic retinopathy using neural networks”, Journal of Theoretical and Apllied Information Technology, vol.88, no.3, pp. 598-606.
11. Lakshmi Prasanna.P, RajeswaraRao. D, 2018, “Text classification using artificial Neural networks”, International journal of Engineering and Technology(UAE), vol.7 no.1.1 Special Issue 1, pp.603-606.
12. Pellakuri V Rao, D.R. Murthy, J.V.R 2016, “ Modeling of supervised adaline neural network learning technique”, proceedings of 2016 2nd International Conference on Contemporary computing and informatics, IC3I 2016,pp.17.
13. Nirmal, K.R.Satyanarayana, K.V.V. 2016, “Issues of Kmeans clustering while migrating to map reduce paradigm with big data: A survey”, International Journal of Electrical and computer engineering, vol. 6 no. 6, pp. 3047-3051.
14. Lakshmi Prasanna . P, RajeswaraRao. D, Meghana. Y, Maithri, K. Dhinesh, T. 2018,” Analysis of supervised classification techniques”, International Journal of Engineering and Technology (UAE), vol. 7 no. 1.1.
15. “Input split frequent pattern tree using mapreduce paradigm in hadoop”, Greeshma . L,G. Pradeepini,2016. Journal of Theoretical and applied information technology, vol. 84,no.2, pp. 260-271.
16. “Adaptive clustering approach to handle multi similarity index for uncertain categorical data streams”, by Kolli, S, Sreedevi . M, 2018. Jounal of Advanced Research in Dynamical and control Systems, vol.10,no. 9 Special Issue, pp. 623-631.
17. Gouse S.M, Krishna Mohan. G, 2018,”A Survey on challenges and future perspectives in big data using cloud-based technologies”, Journal of advanced research in dynamical and control system, vol. 10,no. 9 special issue, pp. 623-631.
18. Buvanesvari. M, Uthayakumar. J, Amudhavel. J. 2017, “Fuzzy based clustering to maximize network lifetime in wireless mobile sensor networks”, Journal of Advanced Research in dynamical and control systems, vol. 9, no. special issue 12, pp.2133-2143.
19. Pratuisha. K,RajeswaraRao. D, Amudhavel. J, Murthy. J.V.R 2017,”Acomprehensive study on artificial neural network techniques for estimation of coronary- artery disease “, Journal of Advanced Research in Dynamical
20. and Control Systems, vol. 9, no. Special issue 12, pp. 1673-1683.
21. Anisha, P.R. VijayaBabu. B, 2018, “EBPS: Effective method for early breast cancer prediction using wisconsin breast cancer dataset”,International Journal of innovative Technology and Exploring Engineering, vol. 8,no. 2S, pp. 205-211.
22. Murthy, K.V.S.S.R. Satyanarayana, K.V.V 2018, “Intrusion detection mechanism with machine learning process – A case study with FMIFSSVM, FLCFSSVM, misuses SVM, anomaly SVM and Bayesian methods”, International Journal of Engineering and Technology (UAE), vol. 7, pp. 277-283.

AUTHORS PROFILE

P. Subhashiniis a student of Department of Computer Science and Engineering studying in KoneruLakshmaiah Education Foundation (K L Deemed to be University) located at Vaddeswaram, Guntur, Andhra Pradesh, India.

K. SatyaVijayais a student of Department of Computer Science and Engineering studying in KoneruLakshmaiah Education Foundation (K L Deemed to be University) located at Vaddeswaram, Guntur, Andhra Pradesh, India.

G.S.K Bhargavis a student of Department of Computer Science and Engineering studying in Konerulakshmaiah Education Foundation (K L Deemed to be University) located at Vaddeswaram, Guntur, Andhra Pradesh, India.



V. Vijay Kumar t of Computer Science and Engineering in KoneruLakshmaiah Education Foundation (KLDeemend to be University). He received degrees in Bachelor of Engineering, Mtech. Currently pursuing Ph.D. Specialized in Data Analytics and Machine Learning. He was experienced in teaching for about 10 years. Research interests are in the field of data analytics and machine learning. He has guided more than 30 Under Graduation students. He published different number of research articles and journals in different platforms. One the experienced professor in K L Deemed to beUniversity.