A NOVEL K-NN CLASSIFICATION APPROACH USING TOPIC MODELLING IN AMINER DATASET

Dr.G.Ayyappan[1] , Dr.C.Nalini[2] , Dr.A.Kumaravel[3]
Associate Professor, Department of Computer Science, Bharath University[1]
Professor, Department of Computer Science, Bharath University[2]
Professor & Head, School of Computing, Bharath Unviersity[3]
ayyappangmca@gmail.com[1], drnalinichidambaram@gmail.com[2], drkumaravel@gmail.com[3]

Abstract - Social network is a structure of human relations and association. It is made up of a organized social actors in a network form. Information has varied number of forms and various purposes for communication. Journals serve as major source of primary information.The topics were divided into categories, such as Algorithm, Data mining, Database, Artificial Intelligence, Clinical, Medical Imaging, Image Processing, Biomedical Informatics, Image Processing, and Telemedicine, which happens to be a little exercise around the topic modeling. The problem of extracting from the huge dataset for author article relationship with appropriate classifier with best accuracy was considered by carrying out the experiment in this chapter.

Keywords: LinearNNSearch, BallTree, Filtered-NeigbhourSearch, Euclidean, Manhattan, and Chebyshev.

I INTRODUCTION

In the background of today's big data, it is an important part of the enterprise activity planning to accurately excavate the interest preference of the user-specific fields from the large data. Nowadays, the emergence of a social network represented by microblog makes a large number of users more willing to use it to share their interest in various fields. Microblog platform will provide a large number of user motion data, which can be used to mine the user's interest preferences in specific areas. Therefore, a lot of user data on the microblog platform can effectively mine the user's interest and bring huge commercial value.

In this paper organizes section one has related works and brief introduction of these fields, In section two represents materials and methods, In section three describes results and discussions and the section four presents conclusion

II MATERIALS AND METHOD

In this section presents materials and methods of this research work. Two components were considered in this section. First one for extracting the themes in the article's abstract by topic modeling and the second one for classifying those identified topics into the domain subject area. Furthermore, the experiment was designed for the classification of the topics focusing on parameter tuning for k-NN classifiers. For this purpose, the dataset collected was named as topic_paper_author in the academic social network data from https://aminer.org/topic_paper_author was shown in Table 1.1 .There are 10 classes in the topic attributes, such as Algorithm, Artificial Intelligence, Biomedical informatics, Clinical, Data Mining, Database, Image Processing, Medical Imaging, Programming, and Telemedicine were shown in Table 1.2.

Table 1.1 Description of Topic_Paper_Author Dataset

S.No.	Attribute	Data type
1	Conference Name	Text
2	Title	Text
3	Year	Numeric
4	Abstract	Text
5	Authors	Text

Table 1.2 Description of 10 classes in topic attributes

S. No.	Class	No. of records
1	Algorithm	3999
2	Artificial Intelligence	630
3	Biomedical informatics	961
4	Clinical	224
5	Data Mining	2496
6	Database	4635
7	Image Processing	4064
8	Medical Imaging	1159
9	Programming	110
10	Telemedicine	97

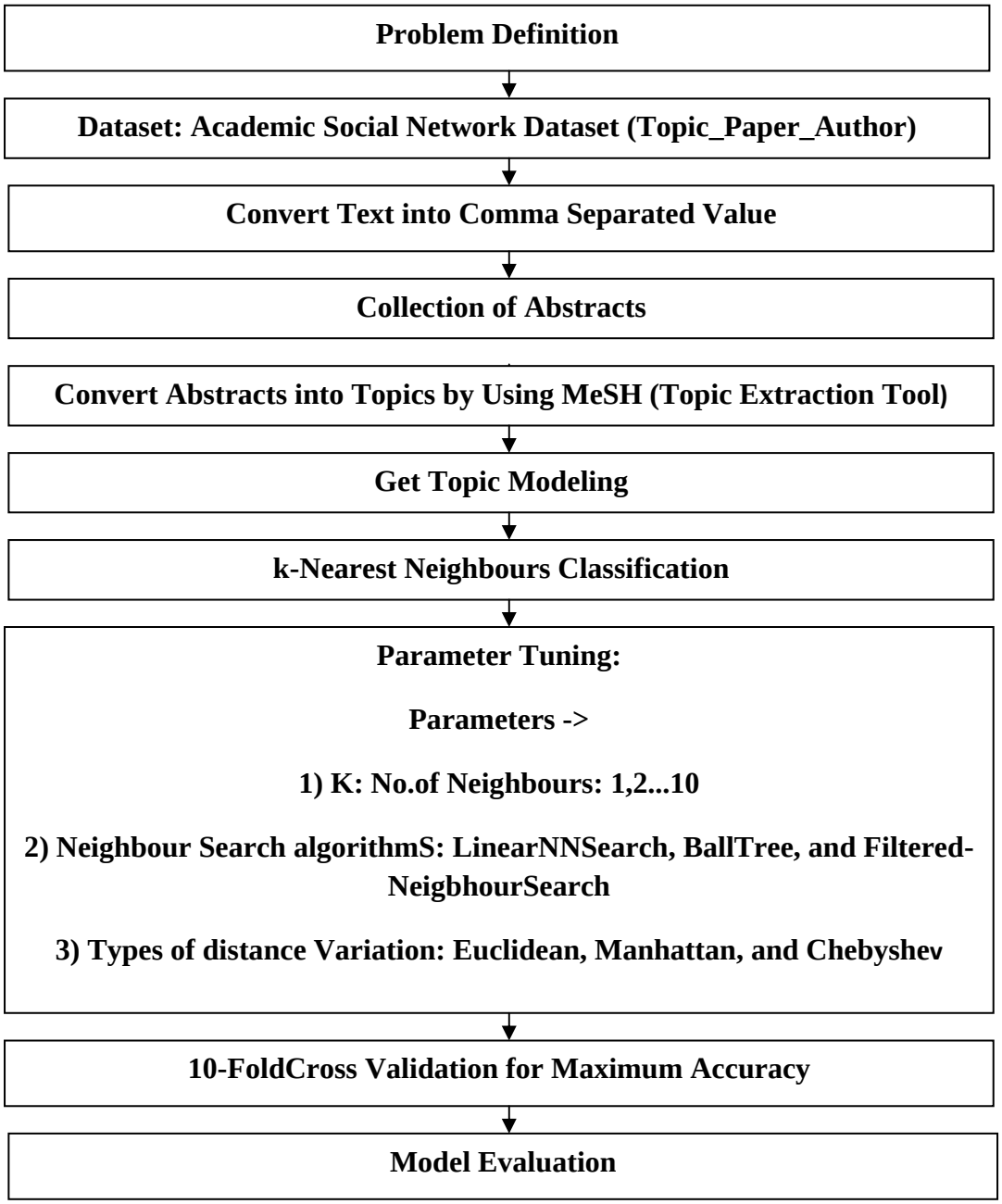

Figure 1.1 Proposed architecture of K-NN.

The above figure represents the flow of the proposed model of the K-NN model using various parameters. Such as, Number of Neighbours, Neighbour search algorithms and the types of distance variations.

The dataset was collected for the purpose of cross domain recommendation. The attributes contain the following segmentation of subject areas.

- **Data Mining:** The data mining conferences were taken from Knowledge Discovery and Data Mining (KDD), Siam International Conference on Data Mining(SDM), The IEEE International Conference on Data Mining series (ICDM), The international Conference on Web Search and Data Mining (WSDM), and European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). It has 6,282 authors and 22,862 co-author relationships.
- **Medical Informatics:** The medical informatics conferences were taken from Journal of Biomedical Informatics, Journal of the American Medical Informatics Association, and Artificial Intelligence in Medicine, IEEE Trans. Med. Imaging, and IEEE Transactions on Information and Technology in Biomedicine. The network has 9,150 authors and 31,851 co-author relationships.
- **Theory:** The theory conferences were taken from The Annual ACM Symposium on Theory of Computing (STOC), The IEEE Annual Symposium on Foundations of Computer Science (FOCS), and ACM-SIAM Symposium on Discrete Algorithms (SODA) conferences. It has 5,449 authors and 27,712 co-author relationships.
- **Visualization:** The visualization were collected from Computer Vision and Pattern Recognition (CVPR),The IEEE International Conference on Computer Vision (ICCV),IEEE Visual Analytics Science and Technology (VAST),The IEEE Transactions on Visualization and Computer Graphics (TVCG),and The IEEE Visualization and Information Visualization conferences included. The co-author network has composed of 5,268 authors and 19,261 co-author relationships.
- **Database:** Association for Computing Machinery's Special Interest Group on Management of Data(SIGMOD), International Conference on Very Large Databases (VLDB), and The IEEE International Conference on Data Engineering(ICDE). It has extracted 7,590 authors, 37,592 co-author relationships.

The first component in this experiment was realized by using the tool MeSH (Medical Subject Headings) (Topic Extraction Tool), which is recommended for extracting the topics for the abstracts available as a column/attribute in the dataset discussed earlier. The topics were divided into categories, such as Algorithm, Data mining, Database, Artificial Intelligence, Clinical, Medical Imaging, Image Processing, Biomedical Informatics, Image Processing, and Telemedicine, which happens to be a little exercise around the topic modeling.

The second component in this experiment was realized by tuning the parameters of the K-NN classifiers for bringing out the better classification accuracy for categorizing the subject catalog for the given dataset, namely “Topic Paper Author” dataset. It contains 18,375 instances and 5 attributes. The parameter tuning was carried out for the three parameters, such as Number of neighbors, Neighbor search algorithms, and Type of distances. The value ranges of parameters have been listed in Table 1.3.

Table 1.3 Parameter tuning details

S.No.	Parameter Name	Parameter Description	Value Range of the Parameters
1	K	Size of the Neighborhood	1,2,…,10
2	Neighbor Searching Algorithm	Method for finding eligible Neighbors	1.LinearNNSearch, 2.BallTree 3.Filtered Neighbor Search
3	Type of distance measure	Denoting the method of calculation of distance between two data points	1.Euclidean 2.Manhattan 3.Chebyshev

It performed the iterations over the value range of the parameters as listed in Table 7.3 using 10-fold cross validation for increasing the accuracy of the model.

III RESULTS AND DISCUSSION

In this section presents the results and discussions of this research work. Various k values in the topic modeling relational data model were applied in this study. The high accuracy values were identified based on Table 1.4. To identify the high accuracy, various methods were used in k-NN, such as LinearNNSearch, BallTree, and Filtered Neighbour Search methods.

Table 1.4 Various K-NN classification accuracies

K-NN	LinearNNSearch		BallTree		FilteredNeigbhourSearch	
	Time Taken to build the Model (In Seconds)	Accuracy	Time Taken to build the Model (In Seconds)	Accuracy	Time Taken to build the Model (In Seconds)	Accuracy
1	0.03	95.37%	0.89	95.37%	0.06	95.37%
2	0.02	95.51%	0.86	95.51%	0	95.51%
3	0.01	95.54%	0.89	95.54%	0	95.54%
4	0.02	95.54%	0.87	95.54%	0	95.54%
5	0	95.54%	0.87	95.54%	0	95.54%
6	0	95.53%	0.89	95.53%	0	95.53%
7	0.2	95.55%	0.87	95.55%	0.01	95.55%
8	0.03	95.55%	0.89	95.55%	0.02	95.55%
9	**0**	**95.55%**	**0.87**	**95.55%**	**0**	**95.55%**
10	**0.01**	**95.55%**	**0.87**	**95.55%**	**0.01**	**95.55%**

The LinearNNSearch method was applied with k-NN =1-10 in the Ibk model in the Weka tool. For k-NN =1, the output was 95.37% with a time consumption of 0.03 Second. For k-NN =2, the output was 95.51% with a time consumption of 0.02 Second. For k-NN =3,4, and 5, the output was the same at 95.54% but the time consumption was 0.02,0, and 0 seconds, respectively. For k-NN = 6, the output was95.53% and the time consumption was 0 second. For k-NN =7 and 8, the output was 95.55% but the time consumption was0.02 and 0.03 seconds. The high accuracy was obtained for k-NN =9 and 10, but the time consumption was 0 and 0.01 seconds, respectively. With regard to the classification of academic social network dataset, for the k-NN method, LinearNNSearch produced the results when tuned for k-NN=1,2,...,10. However, k-NN =9 was recommended because a high accuracy of 95.55% with a time consumption 0 Second was obtained for this social network dataset (Fig. 7.2 and Table 7.4).

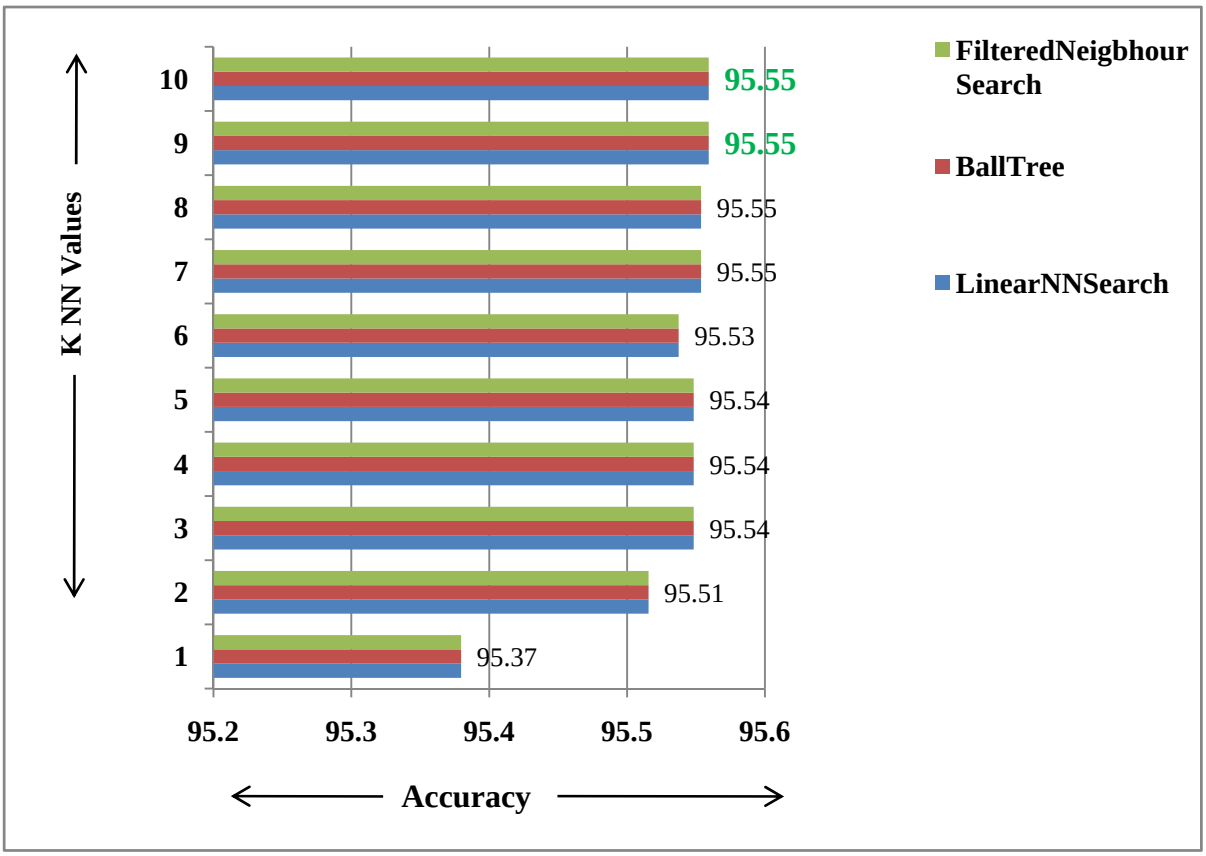

Figure 1.2 Accuracy comparisons among FilteredNeighbourSearch, BallTree, and LinearNNSearch methods.

In the BallTree method, the Ibk model in the Weka tool was applied for k-NN =1-10. For k-NN =1, the output was 95.37% with a time consumption of 0.89 second. For k-NN =2, the output was95.51%with a time consumption of 0.86 second. For k-NN =3,4, and 5,the same output accuracy of 95.54% was obtained, but with a time consumption of 0.89, 0.87,and 0.87 second, respectively. For k-NN = 6, the output was95.53% with a time consumption of 0.89 second. For k-NN =7 and 8, as the same output of 95.55% was obtained but with a time consumption of 0.87 and 0.89 second, respectively. A high accuracy of 95.55% was obtained when k-NN =9 and 10 was applied with a time consumption of 0.87 second.

In the FilteredNeigbhourSearch method, the Ibk model in the Weka tool was applied for k-NN =1-10. For k-NN =1, the output was 95.37% with a time consumption of 0.06 second. For k-NN =2, the output was 95.51% with a time consumption of 0 second. For k-NN =3, 4, and 5, the same output accuracy of 95.54% was obtained with a time consumption of0 second. For k-NN = 6, the output was95.53% with a time consumption of 0 second. For k-NN =7 and 8, as the same output of 95.55% was obtained but with a time consumption of 0.01 and 0.02 second, respectively. A high accuracy of 95.55% was obtained when k-NN =9 and 10 was applied, but with a time consumption of 0 and 0.01 second, respectively.

IV Conclusion

In this experiment shown while the K parameter values 7, 8, 9, and 10 and the accuracy values were above 95.55% for the following three models: LinearNNSearch, BallTree, and FilteredNeighbourSearch. For the K parameter values such as1, 2, 3, 4, 5, and 6, the accuracy values were below 95.55% for these three models are shown in Figure. 7.2.

The problem of extracting from the huge dataset for author article relationship with appropriate classifier with best accuracy was considered by carrying out the experiment in this chapter. In order to identify the best k-NN classifier as the combination (9, FilterNeibourSearch, and Euclidean) as it yields maximum accuracy as evident from the tables and figures enumerated in the previous sections.

REFERENCES

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.
[2] T. Tscharntke, M. E. Hochberg, T. A. Rand, V. H. Resh, and J. Krauss, “Author sequence and credit for contributions in multiauthored publications,” PLoS Biol., vol. 5, no. 1, pp. 0013–0014, 2007.
[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] https://meshb.nlm.nih.gov/MeSHonDemand
[5] https://www.sas.com/en_us/software/university-edition/download-software.html
[6] https://support.sas.com/en/support-home.html
[7] https://www.ultraedit.com
[8] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” IEEE Trans. Neural Networks Learn. Syst., pp. 1–12, 2017.
[9] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.