Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 4837-4850

http://www.ripublication.com

Normalized hamming k-Nearest Neighbor (NHK-nn) Classifier for Document Classification and Numerical Result Analysis

Swatantra Kumar Sahu [1], , Bharat Mishra[2], R. S. Thakur[3] and Neeraj Sahu[4]

[1] *Department of science and Environment, Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, chitrakoot, Satna, Madhya Pradesh, India.*

[2]*Department of science and Environment, Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, chitrakoot, Satna Madhya Pradesh, India.*

[3] *Department of Computer Application, Maulana Azad National Institute of Technology Bhopal, Madhya Pradesh, India.*

[4] *Department of Computer Application, Maulana Azad National Institute of Technology Bhopal, Madhya Pradesh, India.*

Abstract

This paper presents new approach Normalized hamming K-nearest neighbor (NHK-nn) based document classification and numerical results analysis. The proposed classification Normalized hamming K-nearest neighbor (NHK-nn) approach is based on normalized hamming distance. In this paper we have used normalized hamming distance calculations for document classification results. The following steps are used for classification: data collection, data pre-processing, data selection, presentation, analysis, classification process and results. The experimental results are evaluated using MATLAB 7.14. The Experimental Results show proposed approach that is efficient and accurate compare to other classification approach.

Keywords: Normalized hamming k-nearest neighbor, normalized hamming distance, classification and data mining.

1. INTRODUCTION

Document classification is the recent issue in text mining. Document Classification areas are science, technology, social science, biology, economics, medicine and stock market etc.

In last recent years lot of research work has been done decodes some best contributions on Document classification are as follows:

Association Coefficient Measures for Document Clustering [35],An Algorithm for a Selective Nearest Neighbor Decision Rule[2], Hesitant Fuzzy k-Nearest Neighbor Classifier for Document Classification and Numerical Result Analysis[28] , Gradient-Based Learning Applied to Document Recognition[23].

Hesitant Fuzzy Linguistic Term Set Based Laws of Algebra of Sets [29],Condensed Nearest Neighbor Rule[4], Fast Nearest-Neighbor Search in Dissimilarity Spaces[7].

Branch and Bound Algorithm for Computing k-Nearest Neighbors[9], Hesitant Fuzzy Linguistic Term Set Based Document Classification"[30].

Finding Prototypes for Nearest Neighbor Decision Rule[3],An Algorithm for Finding Nearest Neighbors in Constant Average Time[11], Strategies for Efficient Incremental Nearest Neighbor Search[6],Accelerated Template Matching Using Template Trees Grown by Condensation[12], "Document Clustering using message passing between data points"[31].

An Algorithm for Finding Nearest Neighbors[8],A Simple Algorithm for Nearest-Neighbor Search in High Dimension[13], Numerical Result Analysis of Document Classification for Large Data Sets"[32].

A Fast k Nearest Neighbor Finding Algorithm Based on the Ordered Partition[10], Multidimensional Binary Search Trees Used for Associative Searching[14],Discriminant Adaptive Nearest-Neighbor Classification[16], Comparing Images Using Hausdorff Distance[19],Empirical Evaluation of Dissimilarity Measures for Color and Textures[17], A Multiple Feature/Resolution Approach to Hand printed Character/Digit Recognition[18].

Representation and Reconstruction of Handwritten Digits Using Deformable Templates[20],Sparse Representations for Image Decompositions with Occlusions[15],A Note on Binary Template Matching[21], Classification with Non-Metric Distances: Image Retrieval and Class Representation[5],Properties of Binary Vector Dissimilarity Measures[22], Nearest Neighbor Pattern Classification[1],Analysis of Social networking sites using K-mean Clustering algorithm [25].

Computing Vectors Based Document Clustering and Numerical Result Analysis[34], Hesitant Distance Similarity Measures for Document Clustering [27], Classification of

Document Clustering Approaches [24], Hesitant k-Nearest Neighbor (HK-nn) Classifier for Document Classification and Numerical Result Analysis[33],Document Based Classification of Data and Analysis with Clustering Results [26].

The above mentioned work suffers from lack of efficiency and accuracy. The low accuracy is still issue and challenge in the Classification. This motivates us to construct the new method for Classification. New Document Classification method we called normalized hamming K-Nearest Neighbor. Hence we proposed new document classification approach NHK-nn.

The remaining paper is organized as follows: Section-I describe introduction and review of literatures. Section-II describes NHK-nn and K-nn. In Section-III, Methodology of document classification steps are described. In Section-IV, Experimental results are described. In Section-V, results Evaluation and measurement are described. Finally, we concluded and proposed some future directions in Conclusion Section.

2 CALCULATIONS FOR NORMLIZED HAMMING K-NN AND GENERAL K-NN CLASSIFIER

In this calculation we find k- nearest neighbor based on Normalized hamming distance (NHd) and General distance (Gd). Normalized hamming distance and General distance of each p_i to p_j: Table 1 and Table 2. Represent all distance calculated by Hd, Gd. Normalized hamming distance and General distance Cluster Point show in Table 3 and Table 4 with ascending order. This calculation shows normalized hamming distance based accuracy percentages and General distance based accuracy percentages Cluster Point show in Table 5 and Table 6.

For computational model we give tabulation form from Table 1 to Table 6.

Table 1. Normalized hamming distance from Cluster Point

Clusters Points	Normalized hamming distance from Cluster Point P_1 (87,36,77)
P_1 (87,36,77)	0
P_2 (38,7,44)	111
P_3 (120,128,113)	161
P_4 (94,69,8)	109
P_5 (7,98,10)	209
P_6 (105,111,108)	124
P_7 (37,60,88)	85
P_8 (121,123,116)	160

Table 2. General distance from Cluster Point

Clusters Points	General distance from Cluster Point P_1 (87,36,77)
P_1 (87,36,77)	0.00
P_2 (38,7,44)	65.81
P_3 (120,128,113)	104.15
P_4 (94,69,8)	76.80
P_5 (7,98,10)	121.37
P_6 (105,111,108)	83.12
P_7 (37,60,88)	56.54
P_8 (121,123,116)	101.22

3 METHODOLOGY

In the Classification of document different the steps are used. The steps are as follows:

A) Data Collection: In this phase collect relevant documents like e-mail, news, web pages etc. from various heterogeneous sources. These text documents are stored in a variety of formats depending on the nature of the data. The datasets are downloaded from UCI KDD Archive. This is an online repository of large datasets and has wide variety of data types.

B) Classification Method: Initial step is to complete review of literature in the field of data mining. Next step is a detailed survey of data mining and existing Algorithms for Classification. In this area some work done by various researchers. After studying their work, it would be attempted to find the Classification algorithm.

C) Classification Process: Algorithms develop for Classification Process. Classification Process means transform documents into a suitable determined in classes for the Classification task. In Classification Process we performed Different tasks. Optimized classification will also be studied. The real data may be great source for the Classification.

D) Classification Results: In this Experiment we calculate k- nearest neighbor Based on Normalized hamming distance and General distance. Normalized hamming distance and General distance from Cluster Points P_i to P_j calculated and gives ascending order of the normalized hamming distance and General distance for tabulation. Normalized hamming Distance accuracy percentages and General distance accuracy percentages

from Cluster Point show in tabulation. This Experiment show normalized hamming distance based accuracy percentages is efficient and accurate compare General distance based accuracy percentages.

Algorithm 1: This Algorithm obtains Normalized hamming distance of a cluster from each cluster.

Step 1: Input eight clusters points.

Step 2: initialize x_1, y_1 ,$z_1$for cluster point and x_2, y_2,z_2 for each clusters points.

Step 3: Produce and compare normalized hamming distance one by one.

Step 4: find minimum Normalized hamming distance Hd from clusters points say first.

Step 5: arrange all normalized hamming distance in ascending order.

Algorithm 2: This Algorithm obtains General distance of a cluster from each cluster.

Step 1: Input eight clusters points.

Step 2: initialize x_1, y_1 ,$z_1$for cluster point and x_2, y_2,z_2 for each clusters points.

Step 3: Produce and compare General distance one by one.

Step 4: find minimum General distance Gd from clusters points say first.

Step 5: arrange all General distance in ascending order.

4 EXPERIMENTAL RESULTS

In this Experiment we calculate k- nearest neighbor based on Normalized hamming distance and General distance. Normalized hamming distance and General distance from Cluster Points P_1 to P_8 calculated and gives ascending order of the normalized hamming distance and General distance for tabulation describe in Table 3 and Table 4. Normalized hamming Distance accuracy percentages and General distance accuracy percentages from Cluster Point show in Table 5 and Table 6. This Experiment show normalized hamming distance based accuracy percentages is efficient and accurate compare General distance based accuracy percentages.

Table 3. Normalized hamming distance from Cluster Point in ascending order

Clusters Points	Normalized hamming distance from Cluster Point P_1 (87,36,77)
P_1 (87,36,77)	0.00
P_7 (37,60,88)	85
P_4 (94,69,8)	109
P_2 (38,7,44)	111
P_6 (105,111,108)	124
P_8 (121,123,116)	160
P_3 (120,128,113)	161
P_5 (7,98,10)	209

Table 4. General distance from Cluster Point in ascending order

Clusters Points	General distance from Cluster Point P_1 (7,4)
P_1 (87,36,77)	0.00
P_7 (37,60,88)	56.54
P_2 (38,7,44)	65.81
P_4 (94,69,8)	76.80
P_6 (105,111,108)	83.12
P_8 (121,123,116)	101.22
P_3 (120,128,113)	104.15
P_5 (7,98,10)	121.37

Table 5. Normalized hamming Distance accuracy percentages from Cluster Point

Clusters Points	accuracy percentages from Cluster Point %
P_1 (87,36,77)	11.48
P_2 (38,7,44)	22.63
P_3 (120,128,113)	35.59
P_4 (94,69,8)	52.37
P_5 (7,98,10)	61.22
P_6 (105,111,108)	72.29
P_7 (37,60,88)	88.67
P_8 (121,123,116)	96.99

Table 6. General Distance accuracy percentages from Cluster Point

Clusters Points	accuracy percentages from Cluster Point %
P_1 (87,36,77)	9.34
P_2 (38,7,44)	21.34
P_3 (120,128,113)	33.45
P_4 (94,69,8)	48.45
P_5 (7,98,10)	57.45
P_6 (105,111,108)	68.34
P_7 (37,60,88)	78.45
P_8 (121,123,116)	89.45

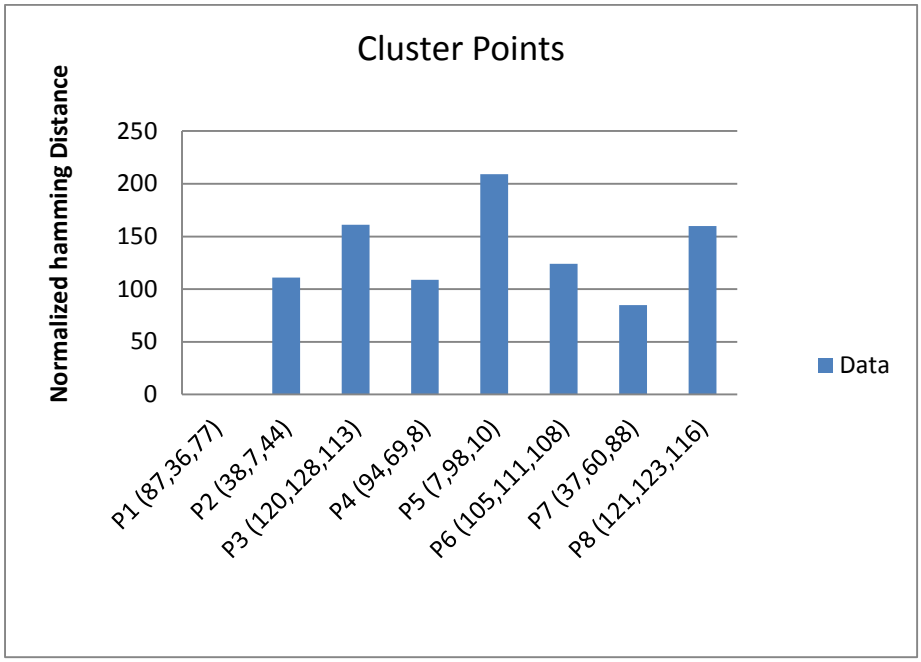

Fig 1: Normalized hamming Distance from Cluster Point.

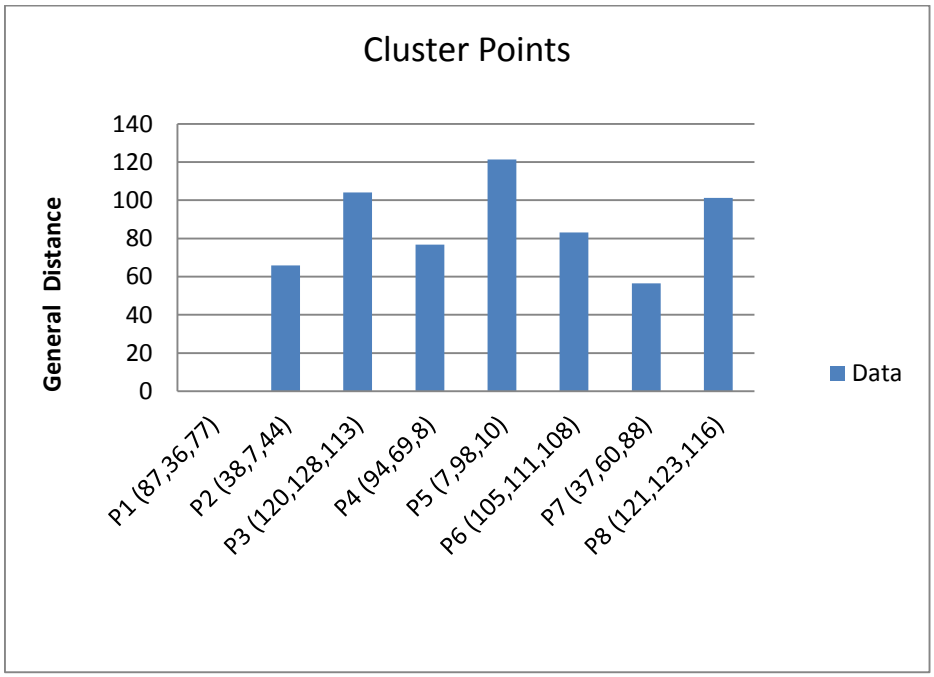

Fig 2: General Distance from Cluster Point

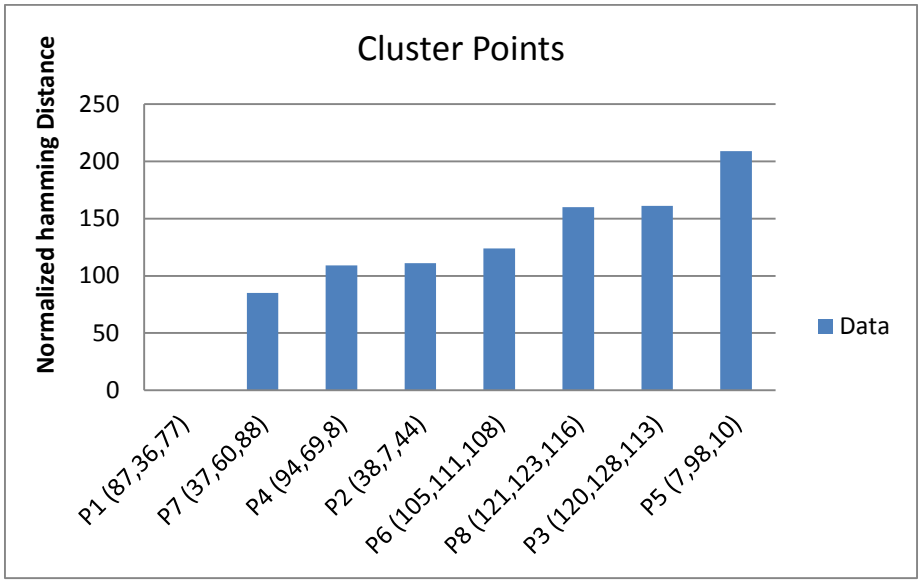

Fig 3: Normalized hamming Distance from Cluster Point in ascending order

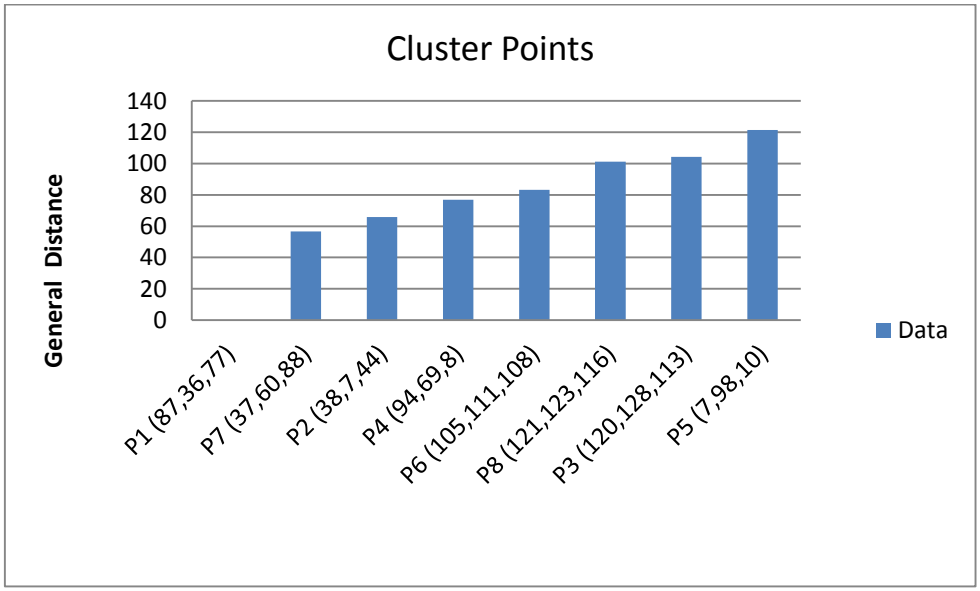

Fig 4: General Distance from Cluster Point in ascending order

The figures 5 and 6 describe document Classification results and Accuracy % of Classification process.

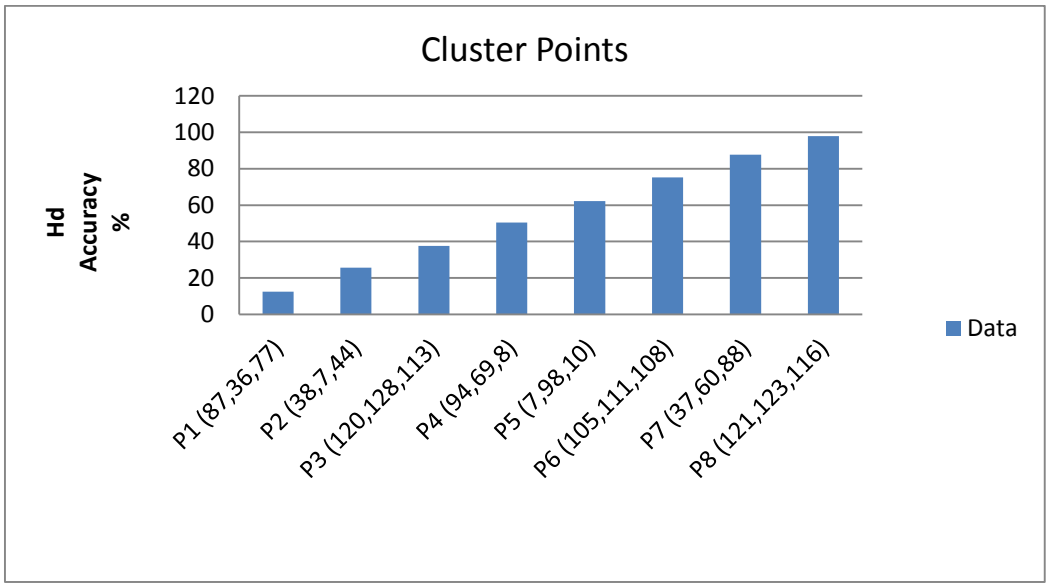

Fig 5: Accuracy % from Cluster Point for Normalized hamming Distance

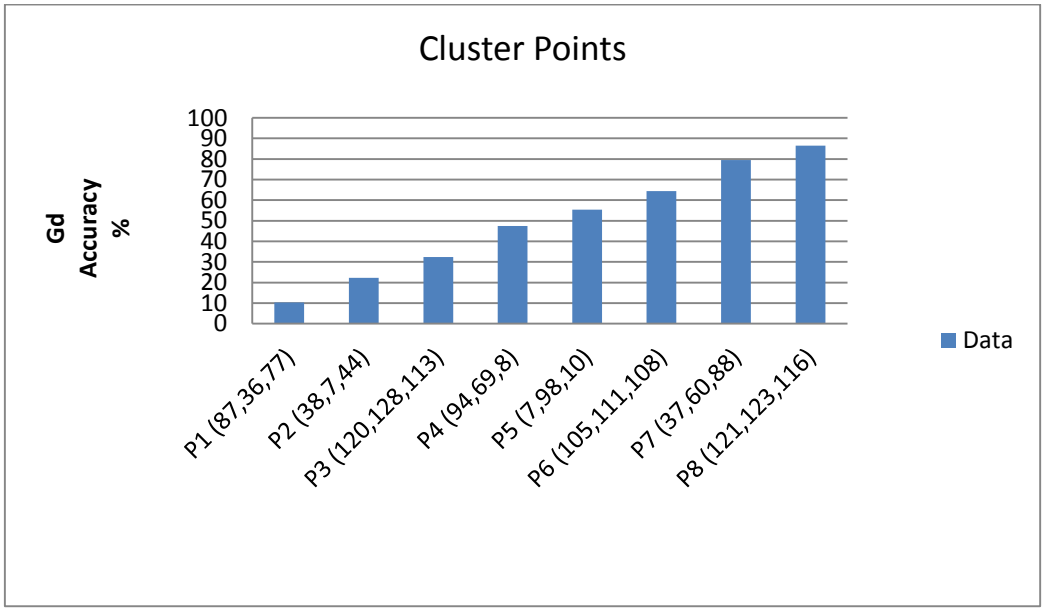

Fig 6: Accuracy % from Cluster Point for General Distance

REFERENCES

[1] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," IEEETrans. Information Theory, vol. 13, pp. 21-27, Jan. 1968.

[2] G.L. Ritter, H.B. Woodruff, S.R. Lowry, and T.L. Isenhour, "An Algorithm for a Selective Nearest Neighbor Decision Rule," IEEE Trans. Information Theory, vol. 21, pp. 665-669, Nov. 1975.

[3] C.L. Chang, "Finding Prototypes for Nearest Neighbor Decision Rule,"IEEE Trans. Computers, vol. 23, no. 11, pp. 1179-1184, Nov. 1974.

[4] P.E. Hart, "Condensed Nearest Neighbor Rule," IEEE Trans. Information Theory, vol. 14, pp. 515-516, May 1968.

[5] D.W. Jacobs and D. Weinshall, "Classification with Non-Metric Distances: Image Retrieval and Class Representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 6, pp. 583-600, June 2000.

[6] A.J. Broder, "Strategies for Efficient Incremental Nearest Neighbor Search," Pattern Recognition, vol. 23, nos. 1/2, pp. 171-178, Nov. 1986.

[7] A. Farago, T. Linder, and G. Lugosi, "Fast Nearest-Neighbor Search in Dissimilarity Spaces," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 15, no. 9, pp. 957-962, Sept. 1993.

[8] J.H. Friedman, F. Baskett, and L.J. Shustek, "An Algorithm for Finding Nearest Neighbors," IEEE Trans. Computers, vol. 24, no. 10, pp. 1000-1006,Oct. 1975.

[9] K. Fukunaga and P.M. Narendra, "A Branch and Bound Algorithm for Computing k-Nearest Neighbors," IEEE Trans. Computers, vol. 24, no. 7,pp. 750-753, July 1975.

[10] B.S. Kim and S.B. Park, "A Fast k Nearest Neighbor Finding Algorithm Based on the Ordered Partition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 761-766, Nov. 1986.

[11] E. Vidal, "An Algorithm for Finding Nearest Neighbors in (Approximately) Constant Average Time," Pattern Recognition Letters, vol. 4, no. 3, pp. 145-157, July 1986.

[12] R.L. Brown, "Accelerated Template Matching Using Template Trees Grown by Condensation," IEEE Trans. Systems, Man, and Cybernetics, vol. 25, no. 3,pp. 523-528, Mar. 1995.

[13] S.A. Nene and S.K. Nayar, "A Simple Algorithm for Nearest-Neighbor Search in High Dimension," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 9, pp. 989-1003, Sept. 1997.

[14] J.L. Bentley, "Multidimensional Binary Search Trees Used for Associative

Searching," Comm. ACM, vol. 18, no. 9, pp. 509-517, Sept. 1975.

[15] M. Donahue, D. Geiger, R. Hummel, and T Liu, "Sparse Representations for Image Decompositions with Occlusions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 7-12, 1996.

[16] T. Hastie and R. Tibshirani, "Discriminant Adaptive Nearest-Neighbor Classification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18,no. 6, pp. 607-615, June 1996.

[17] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi, "Empirical Evaluation of Dissimilarity Measures for Color and Textures," Proc. Int'l Conf. Computer Vision, pp. 1165-1172, 1999.

[18] J.T. Favata and G. Srikantan, "A Multiple Feature/Resolution Approach to Hand printed Character/Digit Recognition," Proc. Int'l J. Imaging Systems and Technology, vol. 7, pp. 304-311, 1996.

[19] D. Hunttenlocher, G. Klanderman, and W. Rucklidge, "Comparing Images Using Hausdorff Distance," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 1, pp. 1-14, Jan. 1997.

[20] A.K. Jain and D. Zongker, "Representation and Reconstruction of Handwritten Digits Using Deformable Templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 12, pp. 1386-1391, Dec. 1997.

[21] J.D. Tubbs, "A Note on Binary Template Matching," Pattern Recognition, vol. 22, no. 4, pp. 359-365, 1989.

[22] B. Zhang and S.N. Srihari, "Properties of Binary Vector Dissimilarity Measures," Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing, Sept. 2003.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proc. IEEE, vol. 81, no. 11, pp. 2278- 2324, Nov. 1998.

[24]. Swatantra kumar Sahu, Neeraj Sahu, G.S. Thakur" Classification of Document Clustering Approaches "International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)", ISSN (ONLINE): 2277 128X Vol-2, Iss-5, May 2012, pp. 509-513.

[25]. D.S Rajput, R.S. Thakur, G.S. Thakur, Neeraj Sahu, "Analysis of Social networking sites using K-mean Clustering algorithm" International Journal of Computer & Communication Technology (IJCCT)", ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 –7449 Vol-3, Iss-3, 2012, pp. 88-92.

[26] Neeraj Sahu, D.S Rajput, R.S. Thakur, G.S. Thakur, "Clustering Based

Classification and Analysis of Data"International Journal of Computer & Communication Technology (IJCCT)", ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 –7449 Vol-3, Iss-3, 2012 pp. 38-41

[27] Neeraj Sahu and G.S. Thakur: "Hesitant Distance Similarity Measures for Document Clustering" IEEE World Congress on Information and Communication Technologies (WICT- 2011) 11 – 14 December 2011, Mumbai ISBN:978-1-4673-0127-5 pp.430 - 438.

[28] Neeraj Sahu, G.S. Thakur" Hesitant Fuzzy k-Nearest Neighbor Classifier for Document Classification and Numerical Result Analysis "International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)", ISSN (ONLINE): 2277 128X Vol-3, Iss-4, April 2013, pp. 920-926.

[29]. Neeraj Sahu, G.S. Thakur" Hesitant Fuzzy Linguistic Term Set Based Laws of Algebra of Sets "International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) ",ISSN (ONLINE): 2277 128X Vol-3, Iss-5, May 2013, pp. 938-945.

[30]. Swatantra Kumar Sahu, Neeraj Sahu, R.S Thakur, G.S Thakur "Hesitant Fuzzy Linguistic Term Set Based Document Classification" IEEE International Conference on Communication System and Network Technologies (CSNT-2013) Gwalior, April 6-8, 2013, ISBN: 978-1-4673-5603-9 ,pp. 586 - 590.

[31]. Neeraj Sahu, Krishna Kumar Mohbey, G.S Thakur "Document Clustering using message passing between data points" IEEE International Conference on Communication System and Network Technologies (CSNT-2013) Gwalior, April 6-8, 2013 ISBN: 978-1-4673-5603-9 ,pp. 591 - 596..

[32]. Swatantra Kumar Sahu, Neeraj Sahu, R.S Thakur, G.S Thakur "Numerical Result Analysis of Document Classification for Large Data Sets"IEEE International Conference on Communication System and Network Technologies (CSNT-2013) Gwalior, April 6-8, 2013 ISBN: 978-1-4673-5603-9 ,pp. 646 - 653.

[33]. Neeraj Sahu, R.S. Thakur ,G.S. Thakur, "Hesitant k-Nearest Neighbor (HK-nn) Classifier for Document Classification and Numerical Result Analysis" Second International Conference on Soft Computing for Problem Solving (SocProS 2012)", Springer 28-30 Dec 2012 Jaipur. ISBN: 978-81-322-1601-8,Advances in Intelligent Systems and Computing, Volume 236, 2014, pp 631-638 .

[34]. Neeraj Sahu, G.S. Thakur, "Computing Vectors Based Document Clustering and Numerical Result Analysis" Second International Conference on Soft Computing for Problem Solving (SocProS 2012) Springer 28-30 Dec 2012 Jaipur. ISBN: 978-81-322-1601-8,Advances in Intelligent Systems and

Computing, Volume 236, 2014, pp 1333-1341.

[35]. Neeraj Sahu, D.S Rajput, Swatandra kumar Sahu, G.S. Thakur, "Association Coefficient Measures for Document Clustering" 2nd International Conference on Computer Science and Information Technology (ICCSIT'2012)Singapore, April 28-29, 2012 pp. 122-126.