Web Document Classification Using Fuzzy K-Nearest Neighbor

Aijazahamed Qazi, R. H. Goudar, P.S.Hiremath

Abstract*: With surge in the number of documents across the internet, increasing the efficiency of any retrieval model is a challenging task. As non-relevant information is retrieved across the internet, increasing the accuracy of any search model is one of the research concerns. Fuzzy classification is broadly applied to address the search issue in search engines. Fuzzy logic provides a methodology to interpret natural language using membership functions. A variant of k-Nearest Neighbor (kNN) called Fuzzy kNN is explored in this paper. This paper provides a comparative analysis of results obtained using kNN and Fuzzy kNN. The Fuzzy kNN results obtained show significant improvement.

***Index Terms*: term, ICF, Fuzzy kNN.**

I. INTRODUCTION

Massive amount of data that is available across the internet can be efficiently classified and retrieved by applying machine learning. Supervised classification models train labeled data to predict test data. kNN is one of the non-parametric algorithms. kNN assigns class label to test data by computing the distance between the nearest neighbors and test data. Zadeh et al. [1] presented the concept of fuzzy logic. It has been extensively used by researchers in many areas of technology. Fuzzy Logic is multivalued in nature. It defines intermediary values using membership functions for computations involving imprecise knowledge. The interpretation of natural language is supported by fuzzy logic. A fuzzy model includes fuzzification and defuzzification with human interpretable rules. The process of fuzzification converts crisp input values to linguistic values within a range [0, 1]. Then membership function is applied to define a fuzzy set for the input. It consists of entities that have partial membership to the fuzzy set. Some of the membership functions used in fuzzy logic are triangular, trapezoidal, gaussian, sigmoid etc. A fuzzy inference engine provides mapping of input to output using fuzzy rules. The process of defuzzification converts the linguistic values into crisp values. The paper is arranged as follows: part 2 of the paper reviews the literature. Part 3 deliberates the preliminaries required. Part 4 explains the proposed approach. The results are discussed in part 5. Part 6 draws the conclusion.

Revised Manuscript Received on September 03, 2019
***** Correspondence Author
Aijazahamed Qazi*, Department of CSE, SDMCET, Dharwad, India.
Dr.R.H.Goudar, Department of CNE, Center for PG Studies Visvesvaraya Technological University, Belgaum, India.
Dr.P.S.Hiremath, Department of MCA, KLE Technological University, Hubli, India.

II. RELATED WORK

Han and Mao [2] presented an approach by applying rough and fuzzy properties of neighbors nearest to the data. The bias was reduced by a membership function and majority class was defined for classification of minority instances. Liu and Chawla [3] presented an improved approach for weighted k-Nearest Neighbor. The classification was performed on imbalanced datasets. Weights defining class confidence was proposed to find the posterior probabilities. Bayesian network and Mixture modeling was applied to calculate the weight of the class confidence. Liu et al. [4] presented a Fuzzy kNN approach for categorical data. The FkNN was coupled and applied to instances of imbalanced data. The proposed approach considered sized membership function and similarity calculation. Ryu et al. [5] introduced a case based selection method using nearest neighbor for the prediction of cross project fault. The proposed approach applied kNN algorithm to learn the local information of the data. The hybrid results obtained indicated improvement in the performance. Patel and Thakur [6] proposed a technique for fuzzy classification of imbalanced data. Experiments conducted showed significant improvement in the performance. Biswas et al. [7] proposed parameter independent fuzzy kNN method. It considered feature weighting technique dependent on class. The value of k was optimized and weights were articulated into a single problem. Maillo et al. [8] introduced a distributed and approximate Fuzzy kNN algorithm using spill tree. The proposed work aimed to address the scalability issues with huge datasets by retaining the classification accuracy. The experiments conducted on the big datasets indicated improvement in the performance.

III. PRELIMINARIES

A. ICF-based term weighting

Term_frequency-Inverse_category_frequency weighting method was introduced by Wang and Zhang [9] to provide lower weight to the terms in several predefined categories. A term's discriminating influence is related to its occurrence within the document and its distribution into various categories.

$$\text{ICF\_based} = \text{Term Frequency} * \log\left(2 * \frac{a}{\max(1,c)} * \frac{|C|}{CF}\right) \quad (1)$$

B. k-Nearest Neighbor

kNN is a lazy learner algorithm without parameter. kNN classifier has been applied in the area of pattern recognition as discussed by Aggarwal et al. [10]. kNN classifier correlates test and training data. The data is represented as a feature vector. For a test vector to be classified, kNN scans the k trained vectors neighboring to the test vector. Then, the test vector is given the class of its nearest neighbors. The Euclidean distance computes the closeness between two points. The Euclidean distance amongst two points, Q = (q_1, q_2, … q_n) and R = (r_1, r_2, … r_n) is,

$$D(Q,R) = \sqrt{\sum_{i=1}^{n}(q_n - r_n)^2} \qquad (2)$$

C. Fuzzy k-Nearest Neighbor

Category membership to the instances of data is assigned by Fuzzy kNN algorithm. It is advantageous for unlabeled query as neighbors are found in prior. The following equations were introduced by Keller et al. [11] towards membership calculation of training data,

$$\text{If } z \in L \text{ and } L = m \text{ then}$$

$$\mu_L(z) = \begin{cases} 0.51 + \left(\frac{n_L}{K}\right) * 0.49 & if\ L = m \\ \left(\frac{n_L}{K}\right) * 0.49 & otherwise \end{cases} \qquad (3)$$

n_L = Nearest neighbors z from category L

$\mu_L(z)$ = Membership of z into class L

And for membership of test data y,

$$\mu_L(y) = \frac{\sum_{i=1}^{K} \mu_{Li}\left(\frac{1}{||y - y_i||^{\frac{2}{p-1}}}\right)}{\sum_{i=1}^{K}\left(\frac{1}{||y - y_i||^{\frac{2}{p-1}}}\right)} \qquad (4)$$

where, p>1 represents an integer and y_i (i=1, 2,…k) is nearest neighbor of y.

IV. PROPOSED TERM WEIGHTING APPROACH

Qazi and Goudar [12] proposed a term weighting method to compute term weight in a document corpus. Initially document preprocessing was carried. Term frequency was obtained for each of the term. Then co-occurrence relation between the terms was calculated. Lin's similarity was computed for the WordNet synonym of the each category and the term. Modified term frequency (semTF) was determined.

$$\text{semTF} = tf_i + \sum_{j=1}^{|N|} C_T(i,j) / |T| + \max_{s_{l_k} \in S_{CL}} \left(SIM_{LIN}(t_i, S_{l_k})\right) \qquad (5)$$

The ICF method was proposed by Wang and Zhang [9] as discussed in preliminaries. The proposed weighting method, semTF_ICF is computed as,

$$\text{semTF\_ICF} = \left(tfi + \sum_{j=1}^{|N|} C_T(I,j) / |T| + \max_{s_{l_k} \in S_{CL}} \left(SIM_{LIN}(t_I, S_{l_k})\right) \right) * \left(\log\left(2 * \frac{a}{\max(1,c)} * \frac{|C|}{CF}\right)\right) \qquad (6)$$

V. EXPERIMENTS

A. Dataset and Evaluation metrics

WebKB benchmark dataset is a group of web pages related to education domain of Washington, Cornell, Texas, and Wisconsin universities. Each university web page is further divided into various categories such as student, department, faculty, course, staff, project and other. The proposed term weighting approach, semTF_ICF is evaluated with the classifiers. The documents are preprocessed. For the classification of the above mentioned dataset, kNN and Fuzzy kNN classifiers are used in the experiments for document classification. The classifiers are discussed in the preliminaries segment of the paper. The performance of the term weighting method is measured using Micro-F1 and Macro-F1 metrics. These metrics are derivative of precision and recall. Micro-F1=2*e*f/(e+f), where e denotes the precision and f denotes the recall of the results. Micro-F1=sum($F1_k$)/m, with k=1,2,…,m and $F1_k$=2*e_k*f_k/(e_k+f_k), where F1 denotes the measure of the k^{th} class, e_k and f_k denote the precision and recall of k^{th} class and m indicates the category count.

Results and Discussion

Performance analysis on WebKB dataset

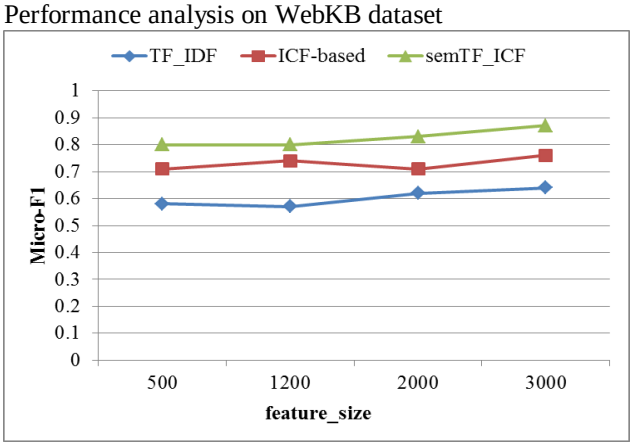

Fig.1.a. Micro-F1 values obtained with kNN (k=14) using three weighting methods on the WebKB dataset.

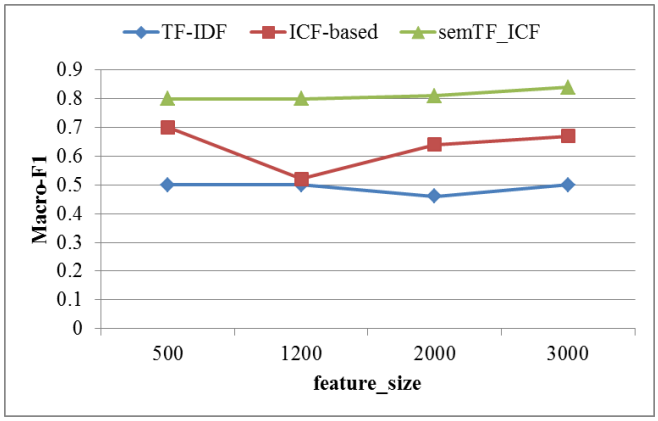

Fig.1.b. Macro-F1 values obtained with kNN (k=14) using three weighting methods on the WebKB dataset.

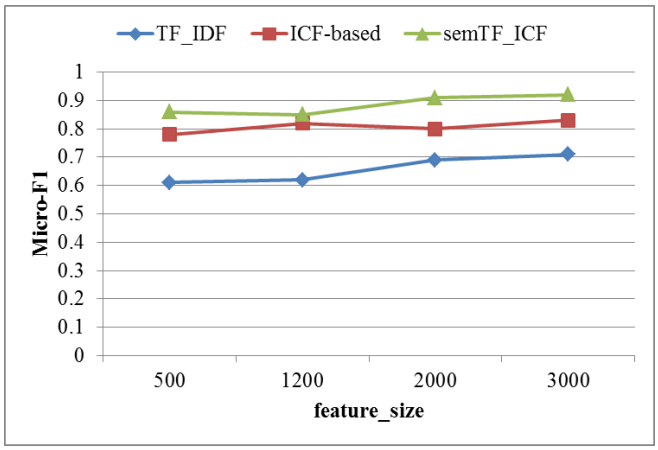

Fig.2.a. Micro-F1 values obtained with Fuzzy kNN using three weighting methods on the WebKB dataset.

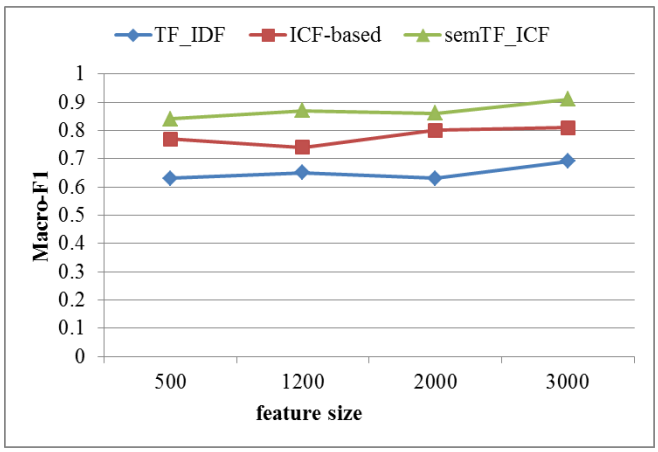

Fig.2.b. Macro-F1 values obtained with Fuzzy kNN using three weighting methods on the WebKB dataset.

The performance of the proposed weighting scheme, semTF_ICF is significantly better than, TF_IDF and ICF-based method on WebKB dataset.

VI. CONCLUSION

Web document classification is an emerging area of research due to the increase in the number of users seeking access of web pages. This paper provides a novel method, semTF_ICF, to weight terms for document classification. A comparative analysis is carried between the results of kNN and Fuzzy kNN classifiers. The results obtained by fuzzy kNN showed significant improvement. As a part of future scope, we intend to examine the method with other machine learning algorithms.

REFERENCES

1. L. A. Zadeh, G. J. Klir, and B. Yuan, Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi A Zadeh, vol. 6. WORLD SCIENTIFIC, 1996.
2. H. Han and B. Mao, "Fuzzy-rough k-nearest neighbor algorithm for imbalanced data sets learning," in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 2010, pp. 1286–1290.
3. W. Liu and S. Chawla, "Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets," in Advances in Knowledge Discovery and Data Mining, vol. 6635, J. Z. Huang, L. Cao, and J. Srivastava, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 345–356.
4. C. Liu, L. Cao, and P. S. Yu, "Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data," in 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 2014, pp. 1122–1129.
5. D. Ryu, J.-I. Jang, and J. Baik, "A Hybrid Instance Selection Using Nearest-Neighbor for Cross-Project Defect Prediction," Journal of Computer Science and Technology, vol. 30, no. 5, pp. 969–980, Sep. 2015.
6. P. Harshita, and T. Singh, "Classification of Imbalanced Data Using a Modified Fuzzy-Neighbor Weighted Approach," International Journal of Intelligent Engineering and Systems, vol. 10, no. 1, pp. 56–64, Feb. 2017.
7. N. Biswas, S. Chakraborty, S. S. Mullick, and S. Das, "A parameter independent fuzzy weighted k -Nearest neighbor classifier," Pattern Recognition Letters, vol. 101, pp. 80–87, Jan. 2018.
8. J. Maillo, J. Luengo, S. Garcia, F. Herrera, and I. Triguero, "A preliminary study on Hybrid Spill-Tree Fuzzy k-Nearest Neighbors for big data classification," in 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, 2018, pp. 1–8.
9. D. Wang and H. Zhang, "Inverse-Category-Frequency based supervised term weighting scheme for text categorization," arXiv:1012.2609 [cs], Dec. 2010.
10. C. C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms," in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 163–222.
11. J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 4, pp. 580–585, Jul. 1985.
12. Qazi, A., & Goudar, R.H.(2019). A Document Classification Framework for Efficient Retrieval. International Journal of Engineering and Advanced Technology, 8(5).

AUTHORS PROFILE

Aijazahamed Qazi, currently working as an Assistant Professor, Dept. of CSE, SDMCET, Dharwad. He has published papers in International Journals and Conferences. His areas of interest include Semantic Web and Information Retrieval.

Dr.R.H.Goudar, currently working as an Associate Professor, Dept. of CNE, Visvesvaraya Technological University, Belagavi. He has 14 years of teaching experience at Professional Institutes across India. He worked as a faculty at International Institute of Information Technology, Pune for 4 years and at Indian National Satellite Master Control Facility, Hassan, India. He has published over 130 papers in International Journals, Book Chapters and Conferences of high repute. His subjects of interest include Semantic Web, Network Security and Wireless Sensor Networks.

Dr.P.S.Hiremath, currently working as a Professor, Dept. of MCA, KLE University, Hubli. He has published several papers in International Journals and Conferences. He has guided many research scholars. His areas of interest include Image Processing, Pattern Recognition and Natural Language Processing.