International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

A Proposed Similarity Measure for Text-Classification

Shitanshu Jain, S. C. Jain, Santosh Vishwakarma

Abstract: ***A Similarity measure is a main process in the text processing method. we have proposed a new method of similarity measure which improved the performance of the K-NN method. The proposed measure extended the accuracy of the text classification method. we have implemented proposed method with Amazon dataset and we observed the effectiveness of proposed similarity measure is increase the accuracy of the similarity between set of documents in a corpus. The end result display performance achieved by way of the proposed measure is higher than that obtained by others.***

Keyword: Similarity Measure, Text Classification, Text Processing, KNN

I.INTRODUCTION

Text processing is very essential and important method of data mining, web search and information retrieval [1], [2]. A record of document is normally represented as a vector in which each aspect indicates the value of the corresponding function within the file. The value of the document function/ feature can be calculated as the number of term occurrences within the document is called term frequency, the factor of the particular term and the number of terms within the set of documents that is relative term frequency or a combination of term-frequency and inverse document frequency (TF-IDF) [3]. Typically the document resulting vector is sparse and dimensionality of a document is very high. I.E., maximum of the function values inside the vector are 0, Such high dimensionality and sparsity can be a severe assignment for similarity degree which is an essential operation in text-processing algorithms

II.TEXT CLASSIFICATION USING K-NN METHOD

k-NN stands for (k-Nearest Neighbor) algorithm which is one of the earliest known technique and one of the simplest classification algorithm used in data mining. Value of "k" is arbitrary and is always given by the user i.e. it is user defined. This algorithm is based on neighbors and the deciding feature for the target class; as to which class an unknown object will belong to is done using votes taken from the neighbors. The factor for predictions is the distance of the object among its neighbors. The unknown object for which we are trying to perform predictive analysis; we analyze all the neighbors of the object.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Shitanshu Jain*, PhD scholar in Amity University Madhya-Pradesh. , India.
Dr. S.C.Jain, Director in the ASET , Amity University, Madhya-Pradesh, India.
Dr. Santosh K. Vishwakarma, Associate Professor, Department of CSE, School of Computing & IT, Manipal University Jaipur, India.

The object will belong to its closest neighbor which is also called as the target class for the object. The measure used for determining neighbors (i.e. how far or how close they are) is done using distance functions such as Euclidean, Manhattan, Hamming distance, etc. The mathematical expression for k-NN distance functions can be given in next section. Classification is a method in which we calculate and measure the distance of each labeled node and unlabeled node of a documents and to check the documents Di belongs or not to class Ca, We determined the similarity and dissimilarity of the two documents (Di and Dj) in the training data set. KNN algorithm complexity of finding the labeled node for an unlabeled node is N X Log A if N Labeled nodes in a class. The file is certainly assigned to a class of its nearest neighbor, if A=1 . The overall performance of the classifier relies upon best on accurate measurement of similarity and dissimilarity parameters.

III.TERM BASED SIMILARITY MEASURES

Similarity of the documents play an important role for the text classification to compute the similarity and dissimilarity among vectors for a document many measures are proposed [9] [10]. Measuring distance is a basic method to calculate the similarity of two documents and it could be used as step of Text-classification [11].
A whole lot of distance measure like cosine similarity measure, Euclidean distance measure, hamming distance measure, Jaccard coefficient index measure, Manhattan distance, Euclidean distance and dice coefficient measure for numeric data-set and many distance metric function are proposed for non numeric data sets.
Euclidean distance: Generally, Euclidean distance metric is not desirable measure of the dimension data extraction application. the Euclidean distance or the distance L2,is the square root of the sum of the squared differences between the elements of the two vectors corresponding. Matching coefficient is a method based on simple vector, it counts the total number of similar terms with which both vectors are non-zero. The coefficient of overlapping regards the two strings as the full match if one is part of another.
Euclidean distance metric is a straight line distance between two points. Euclidean distance or Euclidean metric is the normal distance between two points that measure with a ruler. P1 at (X1, Y1) and P2 at (X2, Y2) in a plane, so it is V ((X1 - X2)² + (Y1 - Y2)²). The formula of distance calculation is given below in equation 1.

$$\sqrt{\sum_{i=1}^{n}(Xi - Yj)^2} \quad (1)$$

Manhattan distance: Manhattan distance is also called bock distance, absolute value distance, L1 distance, boxcar distance and city block distance. Mahnttan Distance Measure calculates the distance between two points which you can travel to get from one point the other point if a path shaped is grid type.

The distance between two elements is a block distance and Block distance is the sum of the differences of the corresponding elements

The distance between two measuring points along axes at right angles. P1 at (X1, Y1) and P2 at (X2, Y2), so it is | X1 - X2 | + | Y1 - y2 | within the plane. The formula of the distance between a point X = (X1, X2) and a point Y = (Y1, Y2) and other points is given below in equation 2.

$$d = \sum_{i=1}^{n} |Xi - Yj| \quad (2)$$

Canberra Distance Metric: Canberra distance measure defined in 19666. It is a distance measure used in data scattered around the origin. Calculation of distance is done by formula of Canberra Distance metric is given in equation 3.

$$d = \sum_{k=1}^{n} \frac{|Yik - Yjk|}{|Yik| + |Yjk|} \quad (3)$$

Cosine similarity Measure: Cosine similarity metric mostly apply in high dimension positive space like text mining and information retrieval. Cosine similarity is a metric used to calculate and measure the cosine of the angle of two non- zero vectors. It computes the similarity of the documents with respect to size of the documents. Mathematically, it computes similarity among two vectors of a multi dimensional space and calculates the cosine of the angle between of them. Cosine similarity metric is mainly used in the positive space, where the result is clearly limited in [0,1].

When the cosine angle is 0 degree (parallel orientation of the vectors) so cosine function result is 1 and for some other cosine angle between 0 degree to less than 90 degree cosine function result is less than 1 and for 90 degree (perpendicular orientation of the vector) cosine function value is 0. As the angles among vectors are less the cosine calculates the similarity value is 1. The formula of calculation of cosine angle is given in equation 4.

$$S(X, Y) = \frac{\sum_{f=1} XiYj}{\sqrt{\sum_{f=1} Xi^2 \ \sum_{f=1} Yj^2}} \quad (4)$$

Chebychev Distance : Chebychave Distance metric is also called maximum measure distance and chess board distance. It calculates the maximum distance or maximum vector space between two points in finite dimensional space. The formula of calculation of Chebychev distance is given in equation 5 when points X=(X1, X2) and Y=(Y1, Y2) and other points are given.

$$Max_i = |Xi + Yi| \quad (5)$$

Hamming Distance: The Hamming Distance is a string metrics for computing the distance between two sequences. Hamming distance is a measure of two strings with same length and calculates the edit distance of number of sequences for which the corresponding symbols are different. S1 and S2 are two strings where S1^ S2 so the Hamming distance between S1 and S2 is evaluated as DH(S1, S2) and measure the places for different X and Y . For example String1 is 10111001 and String 2 is 1001001 so hamming distance is 3 for the given strings. Hamming distance is measure how many bits can be changed in one string to turn into other string. In the above example we need to change 3 bits to turn string1 into string2. The Stings can be represented as String of character instead of the numbers or bits. The formula of calculation of Hamming Distance is given in equation 6

$$HD(S1, S2) = Distance(S1 \hat{} S2) \quad (6)$$

Jaccard Coefficient : The Jaccard Coefficient is a similarity measure which used in computing the comparison of similarity and dissimilarity or diversity of data sets. It is also called the Jaccard index metric. This metric calculate the similarity between sample dataset by using the factor of the size of the intersection and size of the union for sample datasets [14]. A and B are two sample data sets so the jaccard index metric is evaluated as a J(A,B) and measure A and B using formula given in equation 7.

$$I\sqrt{(A, B)} = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

IV. RESULT AND DISCUSSION

In this proposed method, dataset is taken from Amazon e-commerce online shopping website. We have taken 2000 customer reviews for training our system. We performed the data pre processing using Rapid miner tool. It removes the inconsistency and noise from the data. Training of the system is performed using K-NN operator. We applied KNN classifier for classifying customer reviews. in this method we used different similarity measures for the classification and we check the performance of the system. The performance evaluation of system is done using several performance parameters such as accuracy, classification error, Kappa values, and recall and precision values. These performance parameters are analyzed against 2000 records taken for Training Dataset.

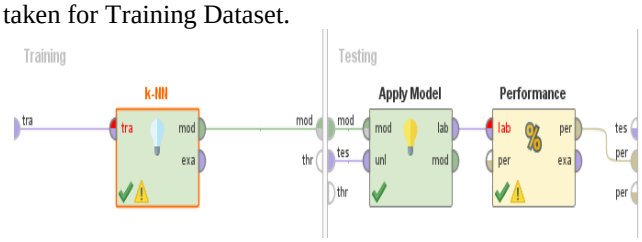

Figure 1: KNN Classifier for Text Classification

We compared our result with different similarity measures for the similar dataset. KNN Classifier generate KNN classification model with different similarity measures and the result of the KNN classifier with different performance parameter is as shown in table.

Table 1: Different Similarity measures with different Performance Parameter

Similarity Measures	Accuracy	Classification Error	Kappa	Precision	Recall
Euclidean Distance	96.38	3.62	0.928	96.13	96.96
Chebychev Distance	96.38	3.62	0.929	95.74	97.14
Correlation Similarity	95.69	4.31	0.915	95.52	95.39
Cosine Similarity	96.43	3.57	0.929	96.16	97.01
Dice Similarity	96.43	3.57	0.929	96.38	97.3
Jaccard Similarity	96.43	3.57	0.929	96.38	97.3
Inner Product Similarity	95.84	4.16	0.918	95.6	96.02
Mahattan Distance	97.03	2.97	0.941	97.08	97.51

In the above mentioned performance comparison table, the proposed method gives highest accuracy with compare to the traditional methods of text classification. The reason is that it has the highest number of true positives thus increasing the accuracy. The KNN method has total of 96.38%, of accuracy while the proposed method has 97.03 % of accuracy. The association between different frequent words is an efficient reason to improve the accuracy of the system. The proposed method also gives the best kappa measure and low classification error. The parameters like weighted mean recall, weighted mean precision and correlation have increased in our proposed approach which is shown in Table 1. Based on the various evaluation parameters we find that the proposed method outperforms in most of the metrics.

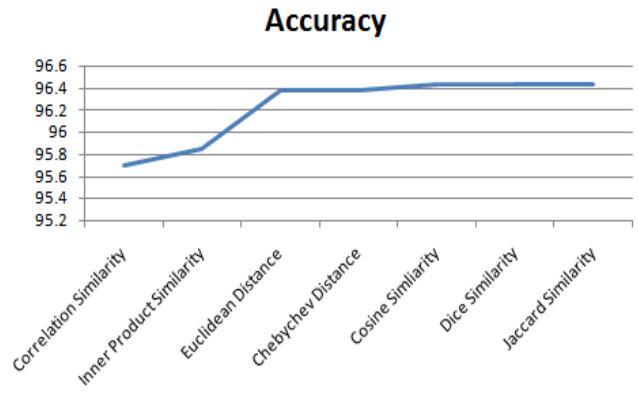

Figure 2: Accuracy parameter for KNN using different similarity measures

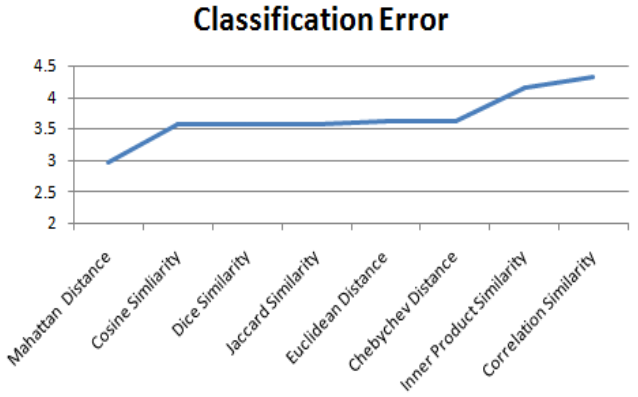

Figure 3: Classification Error parameter for KNN using different similarity measures

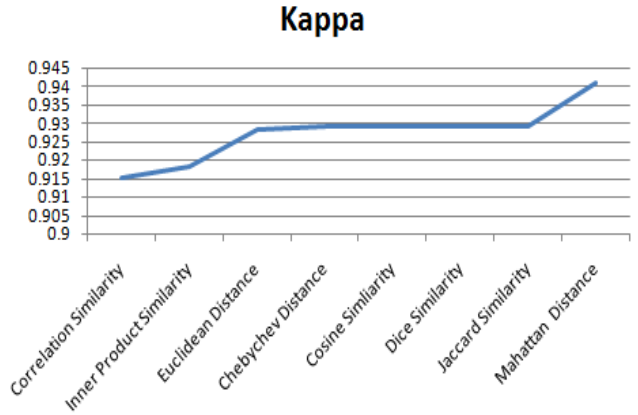

Figure 4: Kappa paramter for KNN using different similarity measures

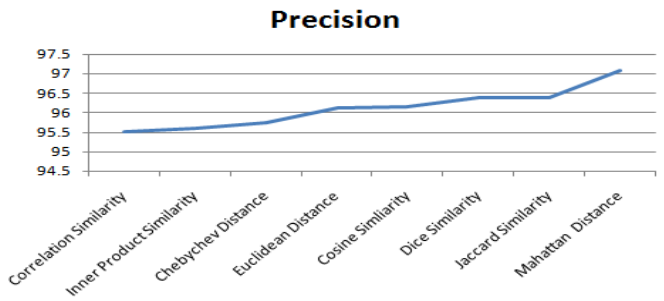

Figure 5: Precision paramter for KNN using different similarity measures

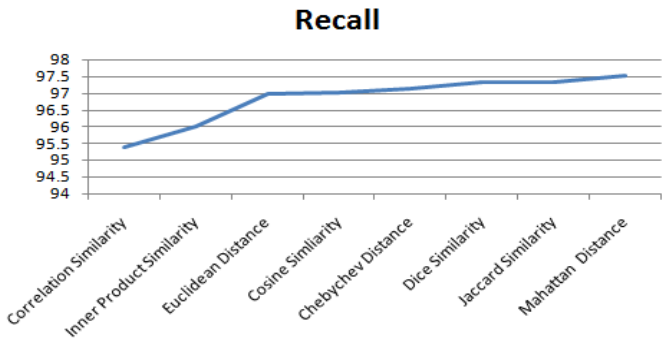

Figure 6: Recall paramter for KNN using different similarity measures

V.CONCLUSION

It has been found that a new way of classifiers can be build which makes KNN Classification techniques applicable to classification tasks and help to solve a number of important problems with the existing classification systems.

The comparison of different measures with KNN classifier methods as mentioned earlier in Table 1, reveals that Proposed method is one of the best classification methods for similarity measures. The higher number of word set from the training dataset reduces the classification error for new document which provide the good results on various parameters.

REFERENCES

1. T. Joachims and F. Sebastiani, "Guest editors' introduction to the special issue on automated text categorization," J. Intell. Inform Syst., vol. 18, no. 2/3, pp. 103–105, 2002.
2. K. Knight, "Mining online text," Commun. ACM, vol. 42, no. 11, pp. 58–61, 1999.
3. [25] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.
4. Christoph Goller, Joachim Löning, Thilo Will and Werner Wolff, 2009, "Automatic Document Classification: A thorough Evaluation of various Methods", "doi=10.1.1.90.966".
5. Megha Gupta, Naveen Aggrawal, 19-20 March 2010, "Classification Techniques Analysis", NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, pp. 128-131.
6. B S Harish, D S Guru and S Manjunath, 2010, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR.
7. Yu Wang and Zheng-Ou Wang, 2007, " A Fast KNN Algorithm for Text Classification", Machine Learning and Cybernetics, International Conference on, Vol. 6, pp. 3436-3441, doi : 10.1109/ICMLC.2007.4370742, Hong Kong, IEEE.

8. Wei Wang, Sujian Li and Chen Wang, 2008, "ICL at NTCIR-7: An Improved KNN Algorithm for Text Categorization", Proceedings of NTCIR-7 Workshop Meeting, December 16–19, Tokyo, Japan.
9. Shraddha Pandit, Suchita Gupta , A Comparative Study On Distance Measuringapproaches For Clustering, International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 1 (2011) pp. 29-31 © White Globe Publications
10. Wikipedia for distance metrics
11. Rui Xu, Donald Wunsch "Survey of Clustering Algorithms" IEEE Transactions on Neural Networks , VOL. 16 NO. 3, MAY 2005.
12. Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem , "An Experiment with Distance Measures for Clustering" , Technical Report: IIIT/TR/2008/132
13. http://en.wikipedia.org/wiki/K-means
14. http://en.wikipedia.org/wiki/Jaccard index

AUTHORS PROFILE

Shitanshu Jain has completed bachelors and master's degree of engineering degree in Computer Science & Engineering from RGPV University, Bhopal. He is currently PhD scholar in Amity University Madhya-Pradesh. His specialization includes Machine learning and data mining algorithms.

Dr. S.C.Jain is working a Director in the ASET ,Amity University, Madhya-Pradesh. He completed his bachelor's and master's degree in Computer Science & Engineering from BITS Pilani, IIT Kharagpur, College of Defence Management. He is a doctorate in the field of Networking.

Dr. Santosh K. Vishwakarma is working as Associate Professor in the department of CSE, School of Computing & IT, Manipal University Jaipur. He completed his bachelor's and master's degree in Computer Science & Engineering. He is a doctorate in the field of Information Retrieval. He holds 15 years of Teaching Experience in reputed Institute. His research interest includes data mining, text mining, predictive analysis