ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SE-K-NN CLASSIFICATION ALGORITHM FOR SEMANTIC INFORMATION RETRIEVAL

Poonam Yadav [*]
[*] D.A.V. College of Engg. & Technology, India

ABSTRACT

Due to the continuous growth of Information retrieval system, text classification is importantly needed to find the category of new information without doing the indexing process again from preliminary. Literature presents different algorithms for classification. Among variety of algorithms, k-NN classification is simple and mostly applied benchmark algorithm for classification. In k-NN classification, similarity matching is an important process which has used different measures. But in most of the measures, semantic way of finding the similarity is completely missing so the way of including semantic keywords based on important words of the documents to the matching process will further improve the matching accuracy. By taking this as motivation, SE measure is developed for the proposed SE-K-NN classification algorithm. In the first step of the proposed algorithm, documents are pre-processed to suit for feature extraction phase and features are computed to build feature database for k-NN algorithm in the second step. In the third step, the devised measure is used for text classification using k-NN algorithm. The experimentation is done with textual database and performance is proved that the proposed algorithm reached of about 75% accuracy as compared with existing algorithm reaches the value of 73%.

KEYWORDS: Information retrieval, k-NN classification, SE measure, accuracy

INTRODUCTION

With the extremely growing amount of information on the Web, textual Information Retrieval systems have turn out to be well-known across the Web community, being used in search engines, online libraries, or local search facilities provided on several websites. A classic search process [2] would occupy users submitting queries, repeatedly in the form of a set of terms, to a retrieval system and receiving a ranked list of results in return. A usual feature of Information Retrieval (IR) systems [3-6] is that if diverse users submit the same query, the system would yield the same list of results. When new information is published as a web page in the web, this page should be indexed properly. But, finding the right category for the new document is a classification problem.

Classifying texts is a fundamental concern in information retrieval and lot of algorithm are utilized in the literature for classification. Classification approaches generally use a training set where all objects are previously related with known class labels. The classification algorithm learns from the training set and constructs a model. The model is used to classify new objects. K-NN [6-10] is a fundamental and simple algorithm for classification widely used by many researchers. Despite the simplicity of the algorithm, it performs very well and is an vital benchmark method. It makes the classification by getting votes of the k-Nearest Neighbors. In k-NN, distance computation is one of the important steps to find the category of the input data. However, in many practical situations, similarities may be preferred over distances. Some standard measures like, cosine measure, Euclidean, Mahalanobis distance and pairwise adaptive similarity are applied for similarity computation.

By analysing these similarity measures, a new measure is devised for text document matching. The proposed SR-measure considers the synonyms of the keywords to match the document in a semantic way. This would improve the performance of the matching even better. The similarity measure developed is then utilized for k-NN classification algorithm. The paper is organized as follows. Section 2 presents the existing Similarity Measures utilized for text processing. Section 3 presents the SE-K-NN classification algorithm for semantic information retrieval. Section 4 presents the results and discussion. Finally, conclusion is given in section 5.

EXISTING SIMILARITY MEASURES FOR TEXT PROCESSING

This section reviews the different similarity measures available in the literature for textual processing. Here, four similarity measures such as Euclidean distance, cosine similarity, pairwise adaptive similarity and SMTP measure are reviewed. Let us consider that d_1 and d_2 are two documents. The document d_1 have $k1$ number of keywords and document d_2 have $k2$ number of keywords. The unique keywords (m) are taken and the frequencies are represented in a vector d_{1_j} . Similarity, the frequencies of unique keywords belonging to d_2 are represented in a vector d_{2_j} . Then, Euclidean distance can be represented as follows

$$d_{Euc}(d_1,d_2)=[(d_1-d_2)\cdot(d_1-d_2)]^{1/2}$$

The cosine similarity of the two documents takes is indicated as,

$$S_{Cos}(d_1,d_2)=\frac{(d_1\cdot d_2)}{(d_1\cdot d_2)^{1/2}\,(d_1\cdot d_2)^{1/2}}$$

The Pairwise adaptive similarity is mathematically formulated as follows,

$$d_{Pair}(d_1\cdot d_2)=\frac{d_{1,K}\cdot d_{2,K}}{(d_{1,K}\cdot d_{1,K})^{1/2}\,(d_{2,K}\cdot d_{2,K})^{1/2}}$$

One of the recent measure, called SMTP measure [1] is defined as follows,

$$S_{SMTP}(d_1,d_2)=\frac{F(d_1,d_2)+\lambda}{1+\lambda}$$

In the above equation, $F(d_1,d_2)$ is defined as ,

$$F(d_1,d_2)=\frac{\sum_{j=1}^{m} N_*\left(d_{1_j},d_{2_j}\right)}{\sum_{j=1}^{m} N_\cup\left(d_{1_j},d_{2_j}\right)}$$

From the above equation, $N_*\left(d_{1_j},d_{2_j}\right)$ and $N_\cup\left(d_{1_j},d_{2_j}\right)$ are defined as follows.

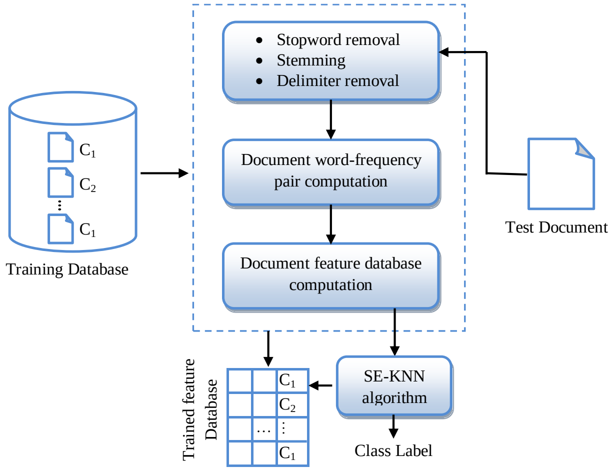
$$N_*\left(d_{1_j},d_{2_j}\right)=\begin{cases}0.5\left(1+\exp\left\{-\left(\frac{\left(d_{1_j},d_{2_j}\right)}{\sigma_j}\right)^2\right\}\right), if\ d_{1_j}d_{2_j}>0\\ 0, if\ d_{1_j}=0\,and\ d_{2_j}=0\\ -\lambda, otherwise,\end{cases}$$

$$N_\cup\left(d_{1_j},d_{2_j}\right)=\begin{cases}0, if\ d_{1_j}=0\,and\ d_{d_j}=0\\ 1, otherwise\end{cases}$$

Drawbacks

Based on the mathematical formulation of four different measures explained above, they have considered the similarity of keywords as main objective to develop the formulae apart from SMTP measure. SMTP measure only considered three different conditions to formulate the similarity degree based on occurrence of keywords in both documents, occurrence of keyword in any one of the document and no occurrence of keywords in both documents. After reviewing these measures, semantic way of finding the similarity is completely missing so the way of including semantic keywords based on important words of the documents to the matching process will further improve the matching accuracy.

Figure 1: Block diagram of the proposed SE-K-NN algorithm



SE-K-NN CLASSIFIATION ALGORITHM FOR SEMANTIC INFORMATION RETRIEVAL

The proposed SE-K-NN classification algorithm is explained in this section. The main contribution of this paper is to devise and implement a new textual measure for finding the similarity between two documents and then, it will be applied to classification of textual documents. Accordingly, the proposed approach of text classification is performed using three different steps. In the first step, documents are pre-processed to suit for feature extraction phase. In the second step, features are computed to build feature database for k-NN algorithm. In the third step, the devised measure is used for text classification using k-NN algorithm. The block diagram of the proposed approach is given in figure 1.

Preprocessing

The input database D of having n number of documents is taken as input for the proposed algorithm. Every document is read out by the programming and important keywords are extracted through the pre-processing steps. In pre-processing, three steps are applied like stop word removal, stemming and delimiter removal. Stop word removal is a process of removing most common words which are not bringing any meaningful for the document if it is considered in separate manner. For example, the words like, 'can', 'could', 'is', 'was' and so on are stopwords. In order to remove these words, stop words list is prepared and directly matching every keyword with stop words list to identify the stop words presented in every document. In the second step, stemming is applied. Stemming is a process to convert all the derived words to its original form. This process is used to bring the derived words to the root form so that further process will be done easily. Then, unwanted delimiter will be identified and removed from the text content. After performing all these processes, document will have only the important keywords which are in the original form without having any symbols or delimiters.

Feature database construction

Feature database FD is constructed by finding the unique keywords from all the documents presented in the original database D. Every element of feature database $FD_{i,j}$ is computed by finding the count of j the unique keyword in i th document. The count can be easily computed by scanning the whole document which are having keywords extracted from the previous step. The feature database which is in the size of $n*m$ where, n is the number of document and m is the unique keywords presented in all the documents. This feature database is used as training features and it is given to the SE-K-NN algorithm for further process.

SE-KNN Algorithm

SE-K-NN algorithm is developed by modifying the standard data mining algorithm, called K-NN algorithm. K-NN algorithm is widely applied for classification due to its simplicity. But, one of the drawbacks of the k-NN algorithm is that it is very sensitive to the similarity computation which affects the performance severely. So, when K-NN algorithm is applied for different application, similarity finding procedure should be done properly based on the applications' need. Accordingly, new measure called SE-measure is developed here for finding the similarity between textual documents when K-NN algorithm is utilized for classification.

Training

The training process is the process of creating the feature database with its class label or ground truth. The previous step was used to build up the database which is then used for the testing process. **Testing:** In the testing phase, a test document is given as input and the corresponding feature vector is build up based on first and second step of the proposed approach. Then, feature vector is matched with every feature vector stored in the feature database using the proposed SE measure. This will given n similarity degree fro all the documents. Based on the minimum value of similarity degree, top k-documents re identified and their corresponding class label is checked. The class label which is having more count will be denoted as class label of the test document.

Proposed SE-Measure

Let us consider that d_1 and d_2 are two documents. The document d_1 have $k1$ number of keywords and document d_2 have $k2$ number of keywords. The unique keywords (m) are taken and frequencies of the keywords are represented in a vector d_{1_j} . Similarity, frequencies of unique keywords belonging to d_2 are represented in a vector d_{2_j} . f_{D_1} is frequency of the keywords in D_1 , f_{D2} is the frequency of keywords in D_2, $\not\neg f_{D_1}$ represents the frequency of keywords in the synonyms set, $\not\neg f_{D_2}$ is the frequency of keywords in the synonyms set.

Figure 2. Algorithm procedure of the proposed SE-K-NN algorithm

```
Input:
Training document, D
Test document, t
A natural number, k
Parameters:
Similarity measure, $SE_{measure}(x,t)$
Pair matrices, U
Output:
Class label of t, r(x)
Begin
    For each x in D do
        Let U ← {}
            Find $SE_{measure}(x,t)$
            Add the pair ( $SE_{measure}(x,t)$ ,  $c(x,t)$ ) to U
            Sort the pairs in U using the first components
            Count the class labels from the first k-elements from U
            Let r(x) be the class with the highest number of occurrence
     End
Return r
End
```

Here, synonyms set are computed by giving the keywords of document to the wordnet ontology. Based on this assumption, the proposed SE-measure is formulated as,

$$\begin{aligned} SE_{measure} = & -\left(P_r\left(D_1, D_2\right)\log P_r\left(D_1, D_2\right)\right) \\ & -\left(P_r\left(D_1, \neg D_2\right)\log P_r\left(D_1, \neg D_2\right)\right) \\ & -\left(P_r\left(\neg D_1, D_2\right)\log P_r\left(\neg D_1, D_2\right)\right) \\ & -\left(P_r\left(\not\neg D_1, \not\neg D_2\right)\log P_r\left(\not\neg D_1, \not\neg D_2\right)\right) \end{aligned}$$

The values of $P_r\left(D_1, D_2\right)$, $P_r\left(D_1, \neg D_2\right)$, $P_r\left(\neg D_1, D_2\right)$ and $P_r\left(\not\neg D_1, \not\neg D_2\right)$ are defined as follows,

$$P_r\left(D_1, D_2\right) = \frac{1}{m}\sum_{i=1}^{m} 2\left(\frac{f_{D_1} + f_{D2}}{\max\left(f_{D_1}, f_{D_2}\right)}\right)$$

$$P_r\left(D_1, \neg D_2\right) = \frac{1}{m}\sum_{i=1}^{m} \frac{f_{D_1}}{\left(f_{D_1} + f_{D_2}\right)}$$

$$P_r\left(\neg D_1, D_2\right) = \frac{1}{m}\sum_{i=1}^{m} \frac{f_{D2}}{\left(f_{D_1} + f_{D_2}\right)}$$

$$P_r\left(\not\neg D_1, \not\neg D_2\right) = \frac{1}{m}\sum_{i=1}^{m} 2\left(\frac{\not\neg f_{D_1} + \not\neg f_{D_2}}{\max\left(\not\neg f_{D_1}, \not\neg f_{D_2}\right)}\right)$$

Procedure

The detailed algorithm procedure of the proposed SE-K-NN algorithm is given in figure 2. From the figure 2, Training document (D), Test document (t) and a natural number (k) are given as user input for the algorithm. The final output (class label) of the test document, t is obtained in the variable, r.

RESULTS AND DISCUSSION

This section presents the experimental results and discussion of the proposed SE-kNN classification algorithm.

Experimental set up

The proposed SE-kNN classification algorithm is implemented with 100 documents having two groups, one is related with sports articles and other one is related with politics' related articles. Every group contains 50 documents and it is given as input to the algorithm. For training, 80% of the documents from each group is taken for building the feature database and remaining 20% of document from every group is used as testing dataset. The obtained classification results are evaluated with the evaluation metrics namely, sensitivity, specificity and accuracy. In order to find these metrics, some of the terms like, True positive, True negative, false negative and false positive is found out based on the following definitions.

Figure 3. Sensitivity plot in between the proposed and existing

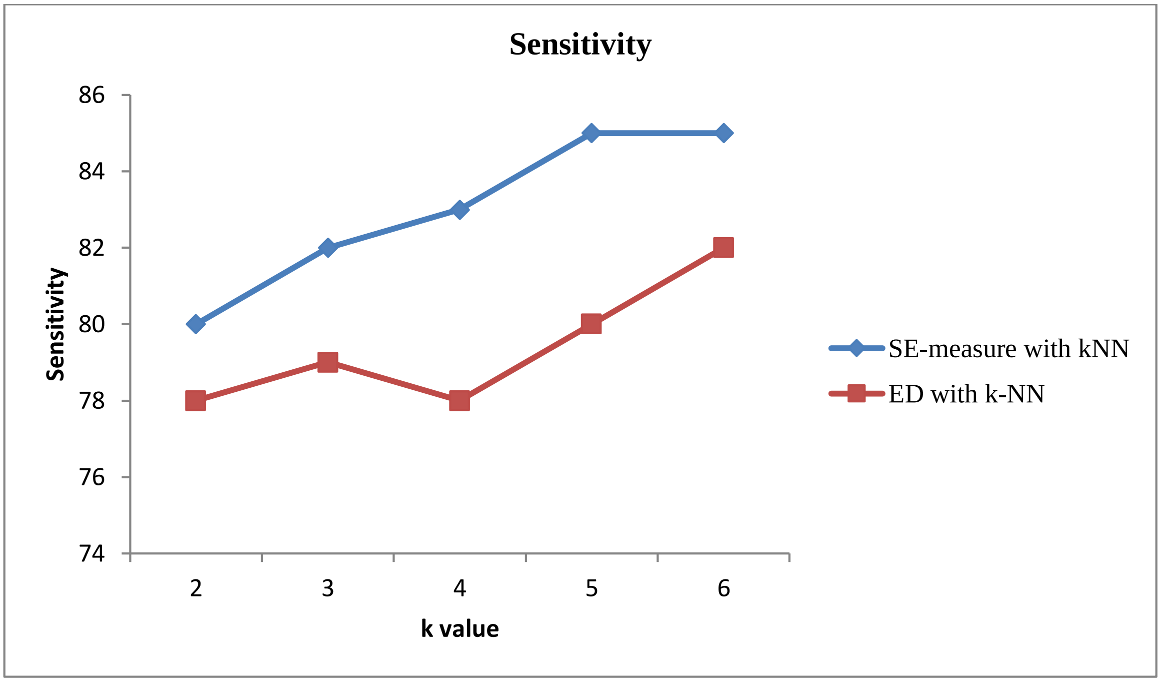


Figure 4. Specificity plot in between the proposed and existing

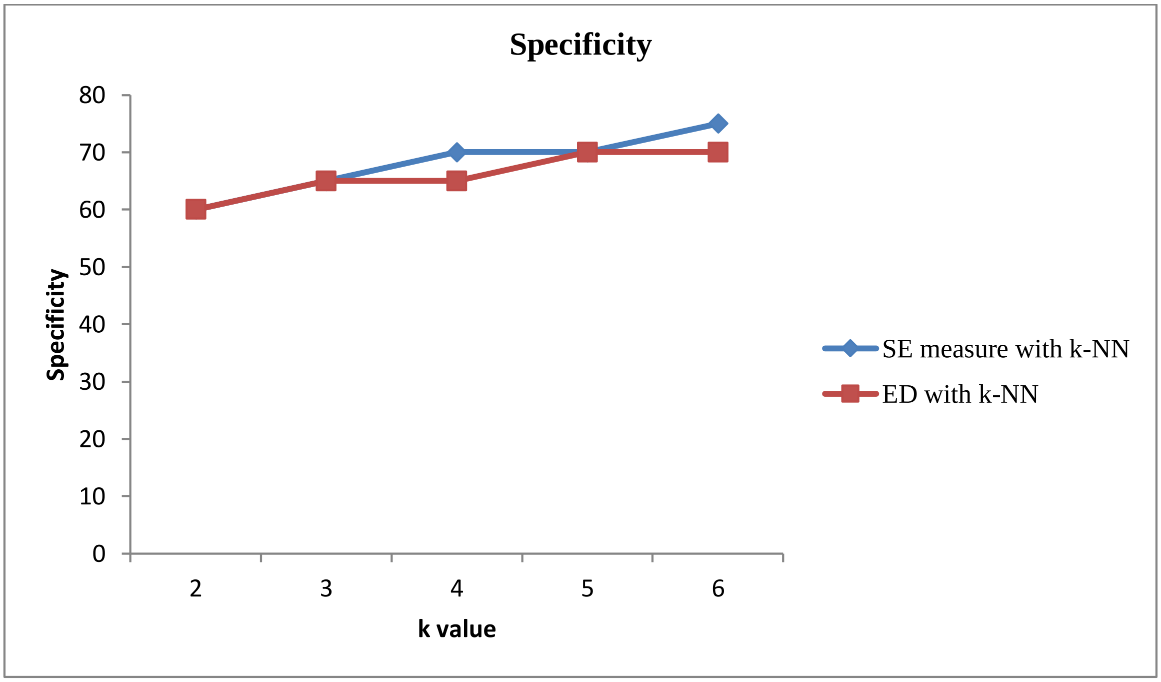
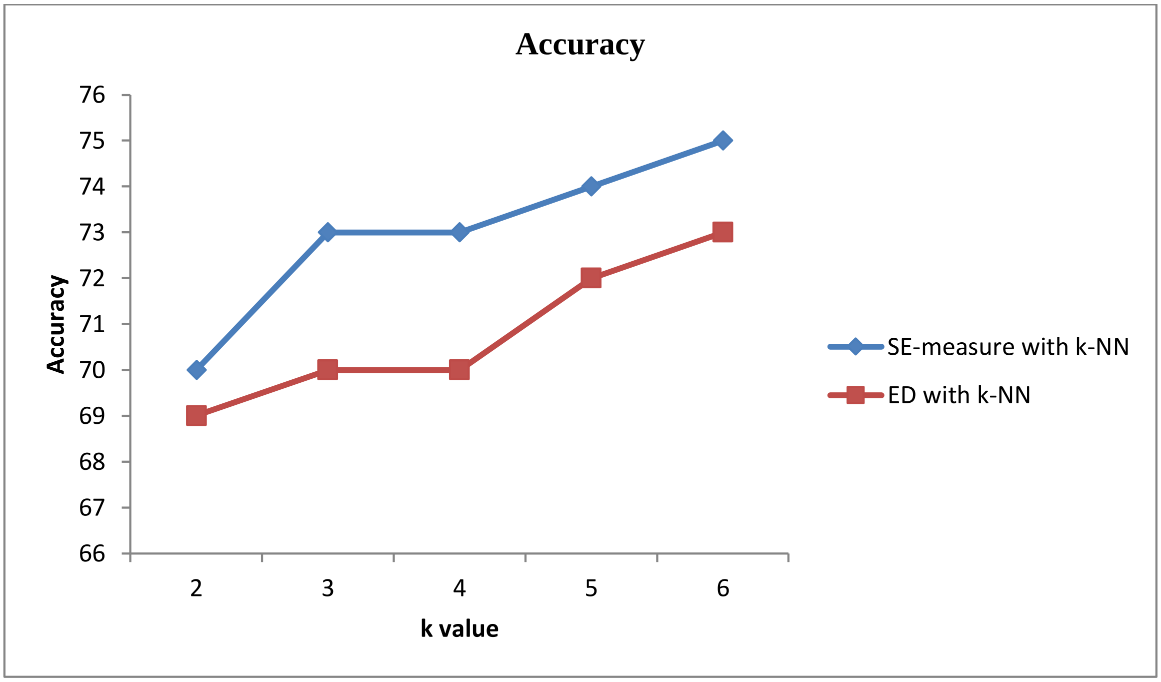


Figure 5. Accuracy plot in between the proposed and existing



$$\text{Sensitivity} = TP/(TP + FN)$$
$$\text{Specificity} = TN/(TN + FP)$$
$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP)$$

where, TP stands for True Positive, TN stands for True Negative, FN stands for False Negative and FP stands for False Positive.

Performance evaluation

The performance plot of the proposed SE-measure with k-NN algorithm and Euclidean distance (ED) with k-NN algorithm is given in figure 3. From the figure, we can easily understand that the proposed algorithm providing good sensitivity for all the different k-values compared with existing algorithm. For k value of six, the proposed algorithm reached of about 85% sensitivity as compared with existing algorithm reaches the value of 82%.

The performance plot of the proposed SE-measure with k-NN algorithm and Euclidean distance (ED) with k-NN algorithm is given in figure 4. From the figure, we can easily understand that the proposed algorithm providing good specificity for all the different k value compared with existing algorithm. For k value of six, the proposed algorithm reached of about 75% specificity as compared with existing algorithm reaches the value of 70%.

The performance plot of the proposed SE-measure with k-NN algorithm and Euclidean distance (ED) with k-NN algorithm is given in figure 5. From the figure, we can easily understand that the proposed algorithm providing good accuracy for all the different k value compared with existing algorithm. For k value of six, the proposed algorithm reached of about 75% accuracy as compared with existing algorithm reaches the value of 73%.

CONCLUSION

A new measure, called SE measure is devised for matching of two textual documents by considering the synonyms of the keywords and occurrence of the keywords in both the documents. Based on the new measure, an algorithm, called SE-K-NN algorithm is developed by extending the traditional k-NN algorithm with SE measure. The semantic way of matching would improve the performance of the matching even better compared with existing algorithm. To prove that, experimentation is done with the existing algorithm using same database and three different evaluation metrics like, sensitivity, specificity and accuracy. According to the experimentation, the proposed approach of text classification obtained sensitivity, specificity and accuracy value about 85%, 70% and 73% respectively as compared with existing algorithm.

ACKNOWLEDGEMENTS

This section should be typed in character size 10pt Times New Roman, Justified.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE transactions on knowledge and data engineering, Vol. 26, No. 7, pp. 1575-1590, July 2014.

[2] Killoran, J.B., "How to Use Search Engine Optimization Techniques to Increase Website Visibility", IEEE Transactions on Professional Communication, vol. 56, no. 1, pp. 50-66, 2013.

[3] Böhm, T, Klas, C.-P. ; Hemmje, M., "ezDL: Collaborative Information Seeking and Retrieval in a Heterogeneous Environment", computer, IEEE, vol. 47, no. 3, pp. 32-37, 2014.

[4] Sumiya, K., Kitayama, D. ; Chandrasiri, N.P., "Inferred Information Retrieval with User Operations on Digital Maps", IEEE Internet Computing, vol. 18, no. 4, pp. 70-73, 2014.

[5] Xiaogang Han, Wei Wei ; Chunyan Miao ; Jian-Ping Mei ; Hengjie Song, "Context-Aware Personal Information Retrieval From Multiple Social Networks", Computational Intelligence Magazine, IEEE, vol. 9, no. 2, 2014.

[6] Junnila, V., Laihonen, T., "Codes for Information Retrieval With Small Uncertainty", IEEE Transactions on Information Theory, vol. 60, no. 2, pp. 976-985, 2014.

[7] Jinn-Min Yang ; Pao-Ta Yu ; Bor-Chen Kuo, "A Nonparametric Feature Extraction and Its Application to Nearest Neighbor Classification for Hyperspectral Image Data", IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 3, pp. 1279-1293, 2010.

[8] Zhao, S. Rui, C.; Zhang, Y., "MICkNN: multi-instance covering kNN algorithm", Tsinghua Science and Technology , IEEE, vol. 18, no. 4, 2013.

[9] Li Ma, Crawford, M.M. ; Jinwen Tian, "Local Manifold Learning-Based k -Nearest-Neighbor for Hyperspectral Image Classification", IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 11, pp. 4099-4109, 2010.

[10] Aslam, Muhammad Waqar, Zhu, Zhechen ; Nandi, Asoke Kumar, "Automatic Modulation Classification Using Combination of Genetic Programming and KNN", IEEE Transactions on Wireless Communications, vol. 11, no. 8, pp. 2742-2750, 2012.

Author Biblography

Dr.Poonam Yadav

Dr. Poonam Yadav obtained B.Tech in Computer Science & Engg. from Kurukshetra University Kurukshetra and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She had been awarded Ph.D in Computer Science & Engg. from NIMS University, Jaipur. She is currently working as Principal in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Her research interests include Information Retrieval, Web based retrieval and Semantic Web etc. Dr. Poonam Yadav is a life time member of Indian Society for Technical Education. Email: poonam.ir@gmail.com