Study of k-NN Evaluation for Text Categorization using Multiple Level learning

Monika
M .Tech Student
Department of Computer Science & Application
M.D. University, Rohtak , Haryana

Rajender Singh Chhillar, PhD
Professor
Department of Computer Science & Application
M.D. University, Rohtak , Haryana

ABSTRACT
Predefined category exists for text categorization. In a document, text may be of any type category like *government, education or health* etc. many methods exist in market invented by researchers for text categorization. One of them is k-NN (k nearest neighbor) algorithm. k play a role to define number of classes for categorization. A training set is generated for each type of category to check its performance than whole text categorized. There is a problem of missing information during training sets. After study recent years invention on k-NN, we find out a solution of this problem. Multiple-Level Learning will improve the performance of k-NN. So in this paper we study about k-NN and propose hybrid algorithm with combination of Multiple-Level Learning and k-NN.

Keyword
Data Mining, Text Classification, k-NN algorithm, Multiple-Level Learning

1. INTRODUCTION
Nearest neighbor search is one of the most popular learning and classification techniques introduced by Fix and Hodges[1], which has been proved to be a simple and powerful recognition algorithm. Cover and Hart [2] showed that the decision rule performs well considering that there is data availability which is no explicit knowledge. In KNN rule, a new pattern is classified into the class with the most members present among the K nearest neighbors, good estimate of Bayes error can be obtain and its probability of error asymptotically ⌈approaches the Bayes error 3]. The traditional KNN text classification has three limitations [4]:

1. **High calculation complexity**: To find out the k nearest neighbor samples, each training samples have similarities with each other that must be calculated. When less number of training samples are available, No longer the KNN classifier is optimal, but if the training set contains a huge amount of samples, more time required by KNN classifier to calculate the similarities. This problem can be solved in 3 ways: reducing the dimensions of the feature space;using smaller data sets; using improved algorithm which can accelerate to [5];

2. **Dependency on the training set:** The classifier is generated only with the training samples and it does not use any additional data. Due to which algorithm depends on the training set excessively; it needs recalculation even if there is a small change on training set;

3. **No weight difference between samples**: All the training samples are treated equally; there is no difference between the samples with small number of data and huge number of data. So it doesn't match the actual phenomenon where the samples have uneven distribution commonly.

A wide variety of methods have been proposed to deal with these problems [6-9]. Another problem is that the classification algorithms will be confused with more number of features. Therefore, feature subset selection is implicitly or explicitly conducted for learning systems [10], [11]. There are two steps in neighborhood classifiers. First an optimal feature subspace is selected, which has a similar discriminating power as the original data, but the number of features is greatly reduced. Then the neighborhood classifier applied. In this paper, we have proposed a novel method based on Rough set theory hybrid with multiple levels learning to select the optimal feature set as discussed in our previous work [11]. Then the proposed MLKNN classifier is analyzed with this reduced feature set.

1.1 Universal Way of Text Classification
This way is used for text classification. Each technique passes through these common steps [12] as shown in fig. 1. After some common steps we apply algorithm which is suitable for us.

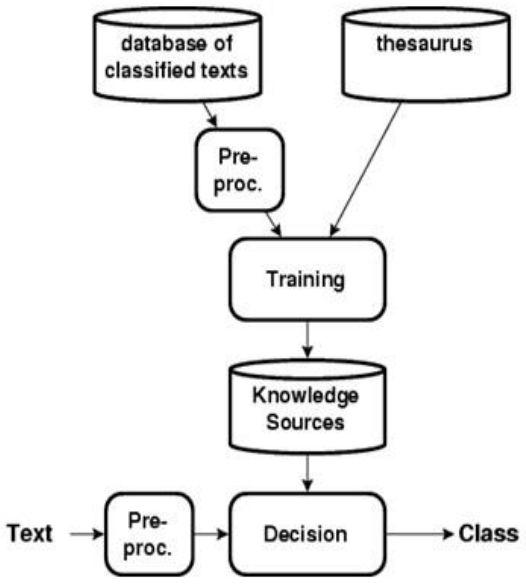


Fig 1. Steps followed for text classification

Training data set is prepared for implementation of text classification process. The flowchart represents decision function where we apply the technique or algorithm to classify data. Training dataset passes through the pre-process in which HTML tags are removed. Size of input data can be reduced in pre – processing data. Like sentence boundary

determination activities involve in it [13]. In knowledge source we can identify important words from documents.

1.2 k-NN Classification Algorithm

k-NN is a case-based learning method, which classify the all data that is in form of training dataset.Due to its laziness feature, it prohibits in many applications such as dynamic web mining for a large repository. Its efficiency can be improve to find some representatives to represent the whole training data for classification, viz.building an inductive learning model from the training dataset and using this model(representatives) for classification. Many algorithms already in market that used by clients such as decision trees or neural networks initially designed to build such a model. One of the evaluation standards for different algorithms is their performance. As k-NN is a simple but effective method for classification and it is convincing as one of the most effective methods on Reuters corpus of newswire stories in text categorization, itmotivates us to build a model for kNN to improve its efficiency whilst preserving its classification accuracy as well.

Looking at Figure 4, a training dataset including 36 data points with two classes{square, circle} is distributed in 2-dimensional data space.

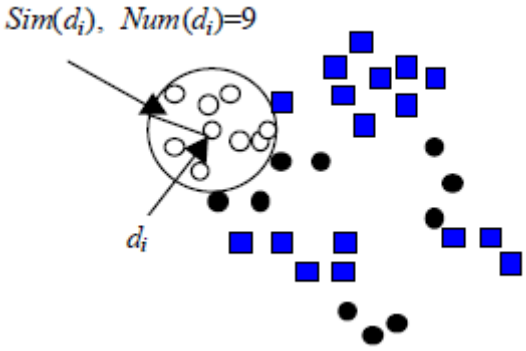


Fig 2. Obtain representative

If we use Euclidean distance as our similarity measure, it is clear that many datapoints with the same class label are close to each other according to distance. The central data point di looking at each local region shown in Figure 4.

1.3 Comparative Results of Algorithms

Effectiveness and share of different group was performed using four experimental sets that were different with each other. As following:

1. Using only features from the General Inquirer (GI).
2. Using only features from WorldNet-Affect (WNA).
3. Combining features from the GI and WNA.
4. Combining all features (including the "other" features comprising of punctuations and emoticons).

Table 1 show the results performed by Naïve Bayes and SVM algorithm on fold of text classification.

Feature	**Navie bayes accuracy**	**SVM accuracy**
GI	71.45%	71.33%
WNA	70.16%	70.58%
GI-WNA	71.70%	73.89%
ALL	72.08%	73.89%

Table 1. Results of text classification

Overall the performance of the SVM classifier was found to be better than that of the Naïve Bayes classifier for this task. The highest accuracy achieved was 73.89%, which surpasses the baseline accuracy of 65.6%. The improvement is statistically significant (on the basis of a t-test, p=0.05). When all features are together then best result was found. There is no effect on result of SVM but it will improve performance of Naïve Bayes.

Table 2 shows result for all feature as following:

MODEL	CLASS	PRECISION	RECALL	F-MEASURE	BASELINE F-MEASURE
CORPUS BASED UNIGRAM	Happiness	0.743	0.377	0.500	0.469
	sadness	0.476	0.341	0.397	0.368
	anger	0.344	0.302	0.321	0.379
	disgonse	0.529	0.320	0.399	0.179
	surprise	0.337	0.243	0.283	0.306
	fear	0.535	0.374	0.441	0.505
	Noemotion	0.394	0.022	0.041	0.579
ROGET'S THEASURUS-FEATURES	Happiness	0.687	0.319	0.436	0.469
	sadness	0.388	0.289	0.331	0.368
	anger	0.400	0.201	0.268	0.379
	disgonse	0.604	0.167	0.264	0.179
	surprise	0.388	0.226	0.286	0.306
	fear	0.672	0.391	0.495	0.506
	Noemotion	0.267	0.013	0.025	0.579
CORPUS BASED UNIGRAM+ RT FEATURE	Happiness	0.690	0.386	0.495	0.469
	sadness	0.368	0.434	0.398	0.368
	anger	0.270	0.346	0.303	0.379
	disgonse	0.387	0.308	0.343	0.179
	surprise	0.256	0.287	0.270	0.306
	fear	0.360	0.426	0.390	0506
	Noemotion	0.471	0.055	0.099	0.579
CORPUS BASED UNIGRAM+RT FEATURES+WNA FEATURES	Happiness	0.698	0.384	0.496	0.469
	sadness	0.361	0.422	0.389	0.368
	anger	0.268	0.358	0.306	0.379
	disgonse	0.402	0.408	0.349	0.179
	fear	0.366	0.426	0.394	0.506

Table 2. Results of fine-grained classification using Naive Bayes

The results from ten-fold cross-validation experiments conducted using the WEKA [14] machine-learning package are shown in Table 3. The performance using the Naïve Bayes classifier was found to be worse than that of SVM.

2. OBJECTIVE OF RESEARCH WORK

We shall follow these objectives as following:

1. Study k-NN algorithm concept in data mining.
2. Calculate parameters value of k-NN like hamming loss, running time, ranking loss, average precision.
3. Use Multiple Label learning with k-NN algorithm.
4. Calculate value of the parameters for this hybrid algorithm.
5. Compare the results.

3. PROPOSED METHODOLOGY

Step1: Select data source on which we apply our proposed algorithm.

Step 2: Implement k-NN algorithm on selected data source and get performance.

Step 3: Now modify k-NN using Multiple Level Learning to develop a hybrid algorithm.

Step 4: Hybrid algorithm ready to perform on same data source.

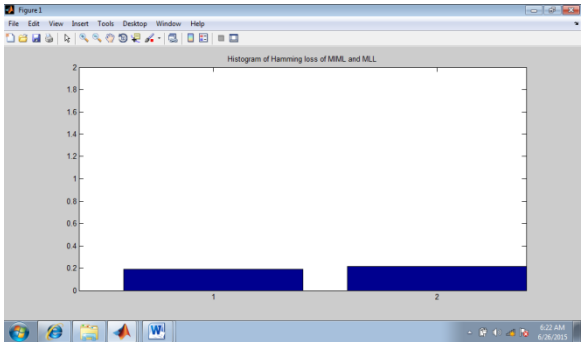
Step 5: performance of hybrid algorithm compared with existing k-NN algorithm.

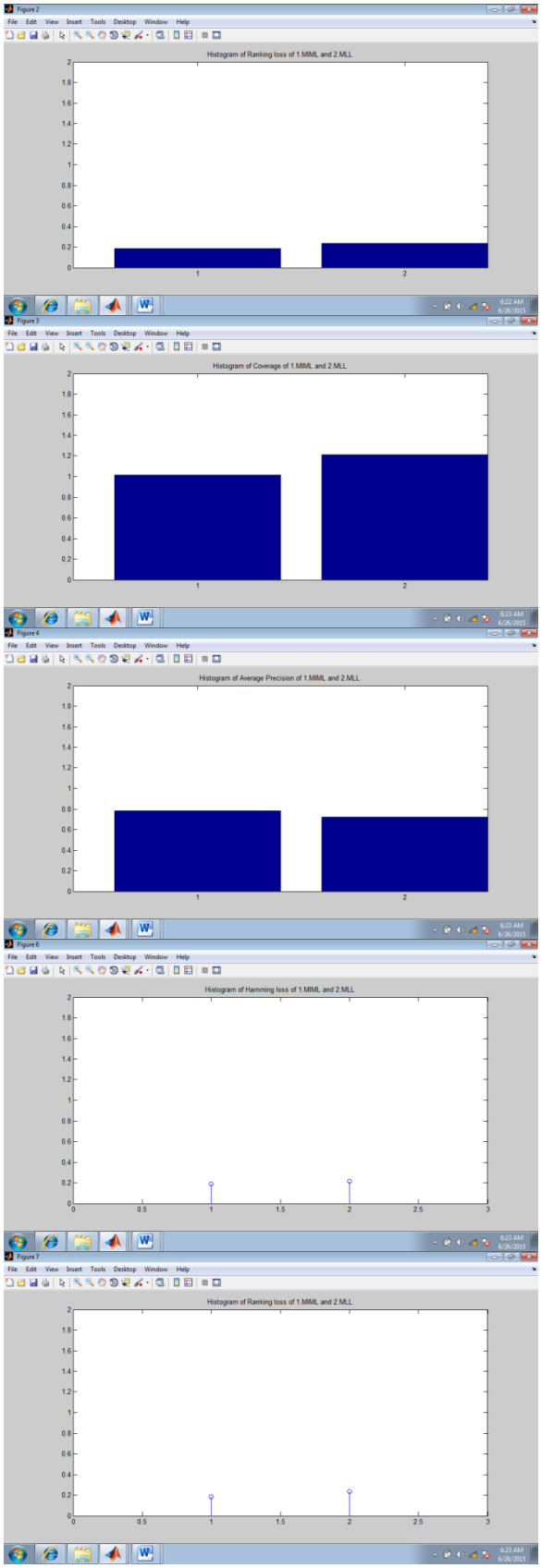
Step 6: Our proposal to develop more efficient algorithm that is named hybrid algorithm in this research work.

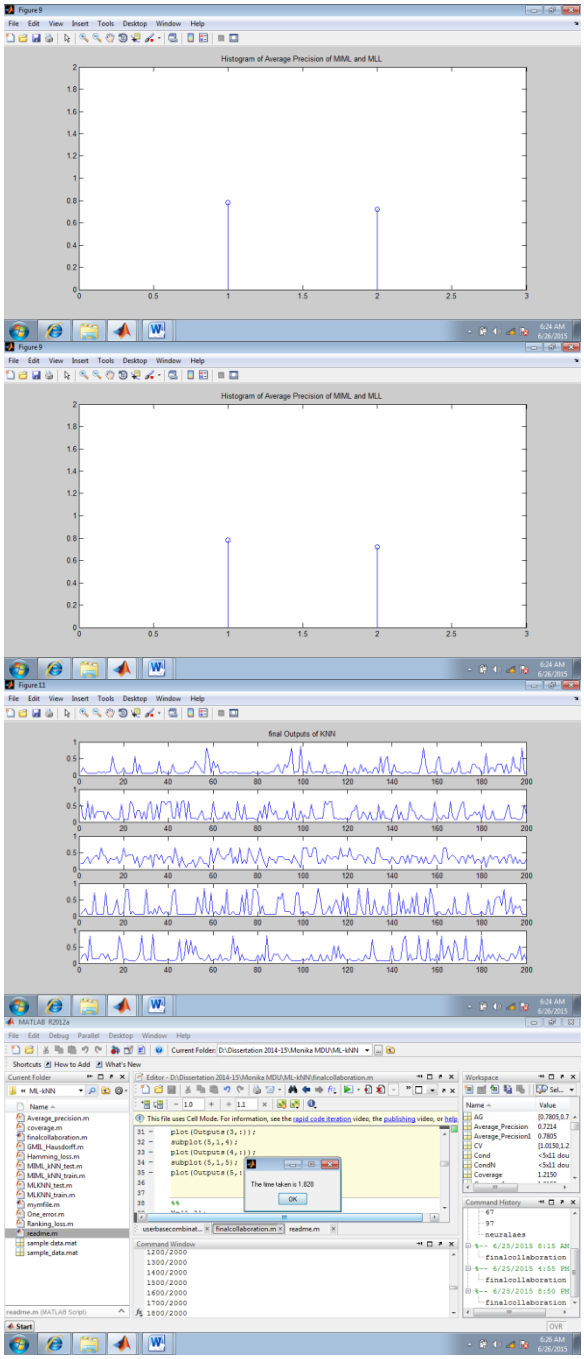
4. CONCLUSION

As we discuss text classification with existing technique SVM, Naïve Bayes, and Neural. SVM perform well in case of number system and other dataset. But this comparison for online text classification, we propose a hybrid algorithm using k-NN with Multiple level learning. The performance of proposed algorithm will be better than these existing algorithms. This proposed algorithm will apply on text file and provide better result. We discuss a concept in objective due to which performance may improve much more as compared to existing. So multiple level learning is helping algorithm for improving performance of k-NN.

5. RESULTS







6. REFERENCES

[1] E. Fix, and J. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties".Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951

[2] T.M. Cover, and P.E. Hart, "Nearest neighbor patternclassification", IEEE Transactions on Information Theory, 13, pp. 21–27, 1967

[3] R.O. Duda, and P.E. Hart, Pattern classification andscene analysis, New York: Wiley, 1973.

[4] W. Yu, and W. Zhengguo, "A fast kNN algorithm for textcategorization", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 3436-3441, 2007

[5] W. Yi, B. Shi, and W. Zhang'ou, "A Fast KNN Algorithm Applied to Web Text Categorization", Journalof The China Society for Scientific and TechnicalInformation, 26(1), pp. 60-64, 2007.

[6] K.G. Anil, "On optimum choice of k in nearest neighborclassification", Computational Statistics and DataAnalysis, 50, pp. 3113–3123, 2006

[7] E. Kushilevitz, R. Ostrovsky, and Y. Rabani, "Efficientsearch for approximate nearest neighbor in highdimensional spaces". SIAM Journal on Computing, 30, pp. 457–474, 2000

[8] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers", Machine Learning, 54, pp. 125–152, 2004, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010 21www.IJCSI.org

[9] C. Zhou, Y. Yan, and Q. Chen, "Improving nearestneighbor classification with cam weighted distance".Pattern Recognition, 39, pp. 635–645, 2006

[10] D.P. Muni, and N.R.D. Pal, "Genetic programming forsimultaneous feature selection and classifier design",IEEE Transactions on Systems Man and Cybernetics PartB – Cybernetics, 36, pp. 106–117, 2006.

[11] J. Neumann, C. Schnorr, and G. Steidl, "Combined SVMbasedfeature selection and classification", Machine Learning, 61, pp. 129–150, 2005

[12] Tanya Taneja, Balraj Sharma, "Text Classification Using PSO & Other Technique" International Journal of Recent Development in Engineering and TechnologyWebsite: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

[13] Zhang W., Yoshida T., and Tang X," Text classification using multi-word features", In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524, 2007.

[14] Witten, I.H. and Frank, E. "Data Mining: Practical Machine Learning Tools and Techniques", (2nd Edition), Morgan Kaufmann, San Francisco, 2005.