International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 29695-29704

http://www.ripublication.com

A Novel Study on Learning Evidence From Distance Metric For KNN Based Text Classification Algorithm

P. Umar Sathic Ali
Assistant Professor, MEASI Institute of Information Technology, Chennai, India
umarsathic@yahoo.com
Dr. C. Jothi Venkateswaran
Associate Professor and Head, Department of Computer Science
Presidency College, Chennai, India

Abstract

Distance metric is most widely used in kNN to measure the similarity estimation between the query pattern and pattern in the training dataset. In this paper we study the traditional distance metric such as Euclidean and we find that these metric may not be appropriate for highly skewed dataset like text categorization. A novel method of learning evidence from multiple distance metric is proposed. Based on DS theory, the evidences learnt from these distance metric are combined for improving the effectiveness of kNN based text classifier. Because the computed neighbours for the given query pattern may be from heterogeneous neighbourhood sources and usually have different influence on predicting the class label. The ensemble of distance metric is tested on three standard benchmark data sets. In all the experiments, robustness of the proposed approach is observed.

Index Terms: Text classification, KNN algorithm, Dempster-Shafer Theory.

1.Introduction

With the rapid growth of development in internet, a huge volume of textual information is prevalent in the form of machine-readable and increases exponentially. In order to leverage the potential from this massive and high dimensional heterogeneous data, text mining and content-based document management task have gradually become the hotspot research field in the research community. Text classification is an important corner stone for information retrieval and text mining, the main task is assigning text document to one or more predefined semantic categories according its content and the labeled training samples [3]. The application of text classification is very broad and has been continuously growing. A typical instance of text classification application is email filtering, where the filter can not

only filter out junk emails, but also distribute emails to the corresponding categories with respect to the content. Text classification technology also play a vital role in commercial web search engines, which provides relevant documents the user is really interested in and thereby filtering unwanted contents. Hence we can infer a conclusion that in the process of managing information related tasks, text classification is a most sought content management method, as it could help users to efficiently organize and retrieve vast amount of information in a hierarchical manner.

At present, the research on text classification has crossed several important milestones in terms of state of art classifiers, standard benchmark datasets and comprehensive empirical evaluation on different algorithms. The classifiers which have been adapted for text classification includes K-nearest neighbor algorithm (kNN), Bayes algorithm, Support Vector Machine algorithm (SVM), decision tree algorithms, neural networks and ensemble of classifiers etc [3].

kNN is one of the most promising classifier for text categorization task and its robustness in improving the effectiveness of text classification has been witnessed by many researchers in the past [3]. However it faces some sort of uncertainties on account of the fact that it rely on single distance metric for similarity estimation without considering the underlying probability distribution. Literature [12] proposed a TBM based kNN classifier to deal with the aspect of ambiguity and imprecise neighbors by associating belief function with each neighbor and combined these belief functions to classify the query pattern. Literature [5] proposed an alternative method for evidence theoretic classification to avoid the need of combining evidence learnt from the neighborhoods. In this method, a single basic belief assignment is constructed for the neighbors. Literature [10] proposed an improved evidence theoretic kNN algorithm for multivariate classification where a multiple distance metric based neighborhood were used and each neighborhood was considered as a piece of evidence to support the class membership of the pattern to be classified. In this paper, we propose a novel method of leaning evidence from distinct distance metric and pooling the evidence associated with the individual metric by the means of DS theory.

2.Review of Text Classification

2.1Problem Description

TC is the problem of approximating the unknown target function $\Phi: D \times C \rightarrow \{T, F\}$ (that defines the way by which documents are to be classified by the human expert) by means of a function $\hat{\Phi}: D \times C \rightarrow \{T,F\}$ called the *classifier*, where $C = \{c_1, \ldots, c_{|C|}\}$ is a predefined set of categories and D is a set of documents. If $\Phi(d_j, c_i) = T$, then d_j is called a positive example of c_i, while if $\Phi(d_j, c_i) = F$ it is called a negative example of c_i. At the outset, this problem looks like a typical classification problem but there are some distinctive characteristics or properties of textual documents which makes the problem as quite challenging to machine learning researchers. These properties are high dimensionality of feature space, statistical sparseness and high level of redundancy.

With respect to task at hand, one may consider TC as either a single-label task (i.e. exactly one $c_i \in C$ must be assigned to each $d_j \in D$), or a multi-label task (i.e. any number $0 \leq n_j \leq |C|$ of categories may be assigned to a document $d_j \in D$). A special case of single-label TC is binary TC, in which, given a category c_i, each $d_j \in D$ must be assigned either to c_i or to its complement $\hat{c}_i$. A *binary classifier* for c_i is then a function $\hat{\Phi}_i : D \rightarrow \{T,F\}$ that approximates the unknown target function $\Phi_i : D \rightarrow \{T,F\}$. A problem of multi-label TC under $C = \{c_1. . . c_{|C|}\}$ is usually handled as $|C|$ independent binary classification problems under $\{c_i, \hat{c}_i\}$, for $i = 1, . . . , |C|$. In this case, a classifier for C is thus actually composed of $|C|$ binary classifiers. This paper deals with evidence learning from distance metrics for binary classifier only.

2.2Text Representation

In order to represent and process the documents by the classifier, each document d_j *is converted* into a compact representation of its content. This can done by a technique called Vector Space Model, borrowed from IR, where a text d_j is typically represented as a vector of words quantified by *weights* $d_j = \langle w_{1j}, . . . , w_{|T|j} \rangle$. Here, T is the controlled *dictionary*, i.e. the set of *terms* (also known as *features*) that occur at least once in at least k documents (in TC: in at least k *training* documents), and $0 \leq w_{kj} \leq 1$ quantifies the term strength of t_k in representing the semantics of document d_j . This transformation process is characterized by what a term is and a method to compute term weights. Regarding term, there are many linguistic features such as word, phrase, sub-word and much more but the most popular choice is to identify terms either with the *words* occurring in the document after eliminating *stop words*, i.e. topic-neutral words such as articles and prepositions, or with their *stems* (i.e. their morphological roots, obtained by applying a stemming algorithm. Regarding a term weights, it may be binary-valued (i.e. $w_{kj} \in \{0, 1\}$) or real-valued (i.e. $0 \leq w_{kj} \leq 1$), depending on the designer of the algorithm to decide whether to include binary input or not. In case of binary weights, weight simply indicates presence/absence of the term in the document. Otherwise, weights are computed by either statistical or probabilistic techniques, the former being the most common option. One of the common statistical criterion to weigh the term is *tf∗idf* where two basic intuitions are playing crucial role, the more frequently t_k occurs in d_j, the more important for d_j it is; the more documents t_k occurs in, the less discriminating it is, i.e. the smaller its contribution is in characterizing the semantics of a document in which it occurs. Weights computed by *tf∗idf* techniques are often normalized so as to contrast the tendency of *tf∗idf* to emphasize long documents.

In TC, another important pre-processing step is to apply a *dimensionality reduction* phase so as to substantially reduce the feature space from T to T' ($T>>T'$) a much smaller feature space in terms of number of features. Dimensionality reduction often takes the form of *feature selection*: each term is scored by means of a scoring function that captures its degree of (positive and sometimes also negative) correlation with c_i, and only the highest scoring terms are used for document representation. Alternatively, dimensionality reduction may take the form of *feature extraction*: a set of artificial terms is generated from the original term set in such a way that the newly

generated terms are both fewer and stochastically more independent from each other than the original ones used to be.

2.3kNN Text Classification Algorithm

The k-Nearest-Neighbor algorithm (*kNN*) is one of the most conceptually simple TC algorithms in the literature. All documents in *D* are considered as vectors in a space with a similarity measure m. To determine whether an unseen document d_i is assigned to a category c_j, the *k* most similar documents to d_i using the measure m are determined, where *k* is a user-adjustable parameter. If the number of these *k* documents that belong to c_j is greater than some predefined threshold, then d_i is assigned to c_j, and otherwise not. This technique has been popularly known as majority voting *kNN* [9].

However, kNN classifier faces serious challenges when patterns of different categories overlap in some regions in the vector space. Moreover, at least two different types of uncertainty may happen namely Ambiguity and ignorance while classifying query pattern [2]. Ambiguity, a type of uncertainty arises in the region of the feature space where there is a strong overlap between classes. In that case, a query pattern x is to be classified will generally be close to several training vectors which belong to different class. Hence this can be viewed as the problem of contemplating evidences.

Ignorance is another type of uncertainty which arises in some situation where the classifier assign a class label to the test pattern on the basis of counting more number of neighbors which are far apart than the fewest number of neighbors which are more close to the query pattern. Thus the classifier actually ignores the close neighbors due to the fact that they are only in small numbers. The application of evidence theory in handling such situations by using multiple distance metrics and thus improving the classifier performance has been proposed in the following section.

3.Learning Evidence From Hetrogenious Neighborhood Based on Theory of Evidence

The most intuitive way to estimate the similarity between two feature vectors is to compute the distance between them using a certain distance metric. In order to estimate the similarity, researchers used the well known metric such as the Euclidean distance. However it has been turned out that this metric may not be suitable for all applications [7]. Moreover, the usage of such metric is justified when the underlying data is drawn from Gaussian distribution [8]. Another important distance metric is Manhattan distance which corresponds to the situation where the data is drawn from Exponential distribution. Unless the underlying data distribution is known in advance, it is highly impossible to infer the corresponding distance metric for the distribution. Most of the prior research work made an implicit assumption that the underlying distribution was either the Gaussian or the Exponential. However such implicit assumption has never always been true. So finding an appropriate distance metric turns out to be a bottleneck if the underlying distribution is neither Gaussian nor Exponential.

Though the most of standard text datasets assume the underlying data distribution is Gaussian, so it is obvious to use L_2 metric in such datasets. However this assumption is inappropriate if the dimensionality of the feature space is very high. Besides, since the document vectors are often extracted for different statistical properties, their distributions may not be the same and different distance metrics may better reflect the distributions for each document vector. Thus, an ensemble of heterogeneous distance metric may be more desirable for estimating the similarity between document vectors.

Motivated by the above discussion and having realised the uncertainty posed by the kNN classifier because of single distance metric, we propose a novel idea of constructing an ensemble of distance metric for kNN based text classifier. Since the underlying data distribution is neither known nor the same for all document vectors, we take in to account more than one distance metric and we create an ensemble of distance metric so that one or more distance metric used in the ensemble may fit the distribution better than any single distance metric. We try to learn evidence from each of the metric used in the ensemble and then combining such evidence by DS theory to improve the performance of the classifier.

Figure 1 shows an input pattern *t* covered by three different neighborhoods. These neighborhoods are obtained by the distance metric L_1, L_2 *and* L_∞ which are then taken as distinct source of evidence in classifying an input pattern *t*.

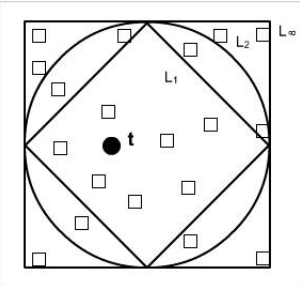

Figure 1: An Example of Pattern *T* In An Overlapped Neighborhoods L_1, L_2 *and* L_∞

The neighborhood provides a source of evidence supporting proposition concerning the class membership of t. Each neighborhood is taken as one part of source of evidence and all neighborhoods – together as a source of evidence – are used to generate a single mass function representing partial support by different neighborhoods. Literature [10] proposed an algorithm which combines such neighborhoods via theory of evidence is detailed below.

An Improved Evidence Theoretic KNN Algorithm

Input: $D = \{(t_i, c_i) : t_i \in V,\ c_i \in W, where\ i = 1,2 \ldots, n\}$ be the training set, where t_i is a vector of d attributes or features whose domain is a relation $V = dom(x_1) \times dom(x_2) \ldots\ldots \times dom(x_d)$ c_i is a class variable whose domain is finite set $W = \{c_1, c_2, \ldots, c_m\}$

Unknown pattern $t = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$

Process:

for i = 1 to h do

Compute the neighborhood E_i with distinct distance metric for the unknown pattern t using the training set D.

end for

for i = 1 to h do

for j = 1 to m do

Set $m[t](E_i, c_j) = \bar{P}(E_i, c_j)/K$

where $\bar{P}(E_i, c_j) = |E_i^{c_j}| / |D|$

$K = \sum_{i=1}^{h} \sum_{c \in W} P(E_i, c)$

end for

end for

for j = 1 to m do

Set $\overline{BetP}(t, c_j) = \sum_{i=1}^{h} m[t](E_i, c_j) / |E_i|$

end for

Set $\overline{BetP}(t) = \sum_{c \in W} \overline{BetP}(t, c)$

for j = 1 to m do

Set $\overline{BetP}(c_j \mid t) = \delta(t) \sum_{i=1}^{h} |E_i^{c_j}| / |E_i|$

where $\delta^{-1} = |D| \times K \times \overline{BetP}(t)$

end for

Set $C = \underbrace{argmax}_{j}\ \overline{BetP}(c_j/t)$

Output: Assign t to class C

4.Experiments and Analysis

The experimental datasets used in this paper are the most popular standard datasets such as Reuters 21578, WebKB and 20 News groups. The details about these data sets are given in Table 1. These datasets are intentionally chosen in order to show the effects of evidence learning on different text domains. In all our experiments, we used information gain as feature selection criteria to reduce the dimensionality of textual documents. For evaluation of text classifiers is concerned, we considered standard information retrieval measures such as precision and recall [3]. Precision is defined as the ratio of correct classification of documents into categories to the total number of attempted classifications. Recall is defined as the ratio of correct

classifications of documents into categories to the total number of labeled data in the testing set. F_1 measure is defined as the harmonic mean of precision and recall. Hence, a good classifier is assumed to have a high F_1 measure, which indicates that classifier performs well with respect to both precision and recall. We choose micro-averaged F_1 measure because of reaching a compromise between Precision and Recall. We present the micro-averaged results for precision, recall, and F_1 measure. Micro-averaging considers the sum of all the true positives, false positives, and false negatives.

Table 1: Summary of The Benchmark Datasets Used In Our Research

Dataset	No. of Documents	Avg. Document Length	No. of Categories	Size	Domain
Reuters 21578 (R8)	7674	193	10	30 MB	News Articles
WebKB	4199	126	4	16 MB	Web Pages
20 News Groups	18821	304	20	56 MB	News Articles

We carried out three experiments to verify the validity of the evidence learnt from the heterogeneous distance metric sources as follows: Experiment 1: This experiment tested the validity of single distance metric by using the traditional KNN classification algorithm. It is observed from the experiments that the traditional KNN algorithm doesn't seem to have satisfactory performance in terms of micro-averaged results and hence it is very necessary to improve the effectiveness of the text classifier. Experiment 2: This experiment is similar to the previous one except that instead of kNN, we considered distance weighted kNN algorithm. The results indicated that the distance weighted kNN has slight improvement over majority voting kNN. This sort of improvement correlates with previous research work reported on this datasets. In both experiments, we used a Euclidian distance metric as a measure of similarity estimation.

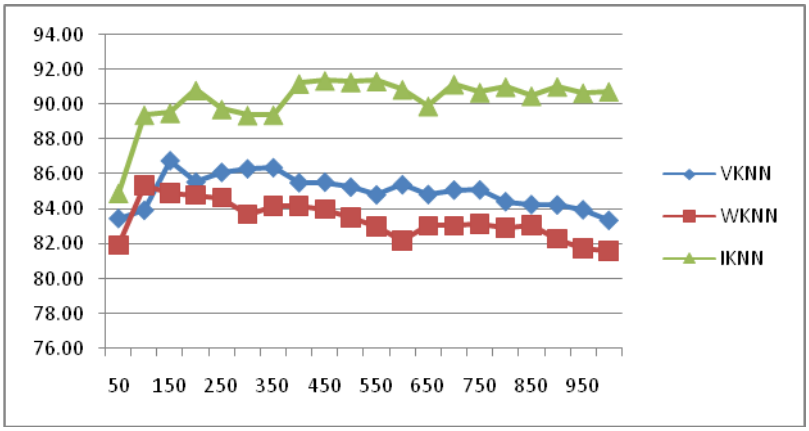

Figure 1: Micro-averaged F1 Measure on Reuters Dataset

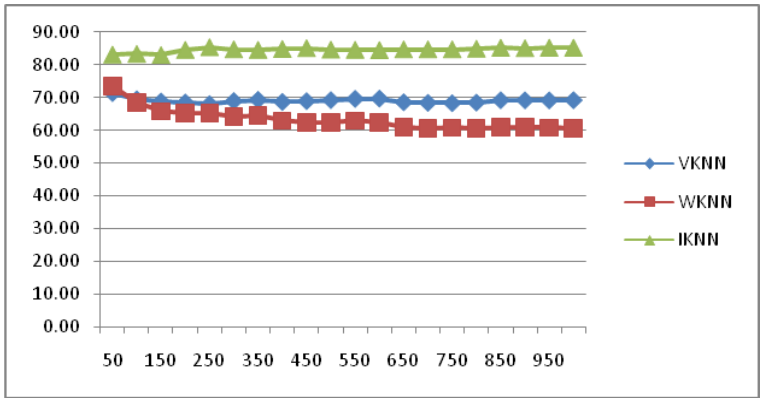

Figure 2: Micro-Averaged F1 Measure on Reuters Dataset

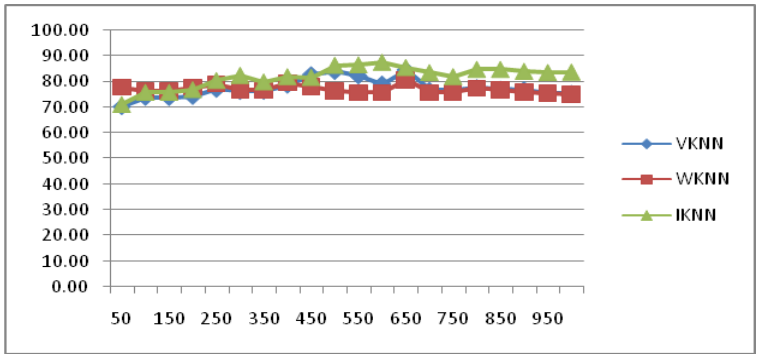

Figure 3: Micro-averaged F1 Measure on WebKB dataset

Experiment 3: We did the last experiment using an improved evident theoretic algorithm. In order to learn evidence, we considered Minkowski family (L_1, L_2 *and* L_∞) of distance metrics as source of evidence. All our experimental results are shown in Figure 1, Figure 2 and Figure 3 respectively. From the

experimental results, it is crystal clear that the micro-averaged F_1 measure has improved greatly when introducing an ensemble of distance metrics for neighborhood calculation for each query pattern to be classified, so this experiment has justified that the validity of combining distance metrics on the basis of the effectiveness the classifier brings out. Figure shows the relationship between F_1 values against number of features chosen from the corresponding datasets. It shows clearly that the values of F_1 measure of IKNN in all three experiments are higher than VKNN and WKNN. It also indicates that the improved evidence theoretic kNN algorithm is better than the traditional kNN algorithms.

5.Conclusion

This paper studied a novel idea of creating an ensemble of multiple distance metrics in order to improve the effectiveness of kNN based text categorization. We tested the validity of the traditional distance metric such as Euclidean by using this metric in association with conventional kNN algorithms and found that these metric are not appropriate for highly skewed dataset like text categorization because of the uncertainty faced by the classifier. We proposed a novel method of learning evidence from multiple distance metric and combining such evidences through DS theory for improving the effectiveness of kNN based text classifier. We performed three experiments using conventional kNN, distance weighted kNN and improved evident theoretic kNN on three standard benchmark data sets. Experimental result on these dataset proves that the ensemble of distance metric has significant influence on improving the effectiveness of kNN based text classifier.

6.References

[1] E. Chavez, G. Navarro, R. Baeza-Yates, J.L. Marroquin. Searching in Metric Spaces, *ACM Computer Surveys*, Vol. 33:3, pp. 273-32, 2011.

[2] E. H¨ullermeier, 'On the Representation and Combination of Evidence in Instance-Based Learning',ECAI–2000, pp. 50–54, 2000.

[3] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[4] G. Shafer, "A mathematical theory of evidence" Princeton University Press,Princeton, New Jersey (1976).

[5] H. Wang and David Bell."Extended k-Nearest Neighbours based on Evidence Theory", The Computer Journal, vol 47, pp 662–672, 2004.

[6] J. Amores, N. Sebe, P. Radeva, "Boosting the distance estimation: application to the K-nearest neighbor classifier," *Pattern Recognition Letters*, Feb. 2006.

[7] M. Zakai "General distance criteria," *IEEE Trans. On Information Theory*, pp. 94-95, January 1964.

[8] N. Sebe and M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1132-1143, Oct. 2000.

[9] P. Cunningham, and Sarah Jane Delany,: k-Nearest Neighbour Classifiers, Technical Report UCD-CSI-2007-4, 2007.

[10] P.Umar Sathic Ali and Jothi Venkateswaran.C: An Improved Evidence Theoretic *k*-NN classifier based on Theory of Evidence. *International Journal of Computer Applications*,15 (5), 37-41. February 2011.

[11] S.A. Dudani. The distance-weighted k-nearest neighbor rule. *IEEE Trans.Syst. Man Cyber.*, 6, 325–327, 1976.

[12] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory". IEEE Transactions on Systems, Man and Cybernetics:, 25, 804–813,1995.

[13] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, IT-13, 21–27, 1967.