International Journal of Mathematical Archive-3(12), 2012, 4781-4788
Available online through www.ijma.info ISSN 2229 – 5046

# INFORMATION THEORETIC METHOD OF FEATURE SELECTION FOR TEXT CATEGORIZATION

**Parmil Kumar* & Sunny Babber**
*Department of Statistics, University of Jammu, Jammu, India*

*(Received on: 09-11-12; Revised & Accepted on: 20-12-12)*

***ABSTRACT***

*With the rapid spread of the Internet and the increase in on-line information, the technology for automatically classifying huge amounts of diverse text information has come to play a very important role in today's world. In the 1990s, the performance of computers improved sharply and it became possible to handle large quantities of text data. This led to the use of the machine learning approach, which is a method of creating classifiers automatically from the text data given in a category label. This approach provides excellent accuracy, reduces labor, and ensures conservative use of resources. In this communication, we discussed that Feature selection plays an important role in Text Categorization (Yiming Yang, Jan O. Pedersen, 1997).We have also deliberated on Automatic feature selection methods such as document frequency thresholding (DF), Information Gain (IG), Mutual Information (MI) and Pointwise Mutual Information (PMI) which are commonly applied in text categorization.*

## 1. INTRODUCTION

With the rapid growth of information available on web in digital form text categorization has become one of the key techniques for handling and organizing the huge text data. Text categorization is the technique of automatically assigning to predefined categories (free text documents), As more and more information is available on online resources, so need arises of good indexing and summation of document contents for effective retrieval. In recent years, a growing number of statistical classification methods and machine learning techniques have been applied in this field. A major difficulty of text categorization process is the high dimensionality of the feature space, which can be tens of thousands, even for a moderate sized text collection. This is prohibitively high for many learning algorithms.In this section we have defined the commonly used terms in the present paper.

The entropy of a random variable is considered as a measure of the uncertainty of the random variable, it is a measure of the amount of information required on average to describe the random variable. The entropy H(X) of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in X} p(x)\, log\, p(x)\,. \tag{1.1}$$

The joint entropy $H(X,Y)$ of a pair of discrete random variables $(X,Y)$ with a joint probability distribution $p(x,y)$ is defined as

$$\mathrm{H(X,Y)} = -\sum_{x\in X}\sum_{y\in Y} \mathrm{p(x,y)\ log\ p(x,y)} \tag{1.2}$$

If $(X,Y) \sim p(x,y)$, then the conditional entropy $H(Y|X)$ is defined as

$$\mathrm{H\,(Y|X)} = \sum_{x\epsilon X} \mathrm{p(x)H(Y|X=x)} \tag{1.3}$$

Next, we introduce two concepts related to each other viz. Relative entropy and Mutual information.

The relative entropy or Kullback–Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\mathrm{D\,(p||q)} = \sum_{x\in X} \mathrm{p(x)log}\frac{\mathrm{p(x)}}{\mathrm{q(x)}} = = \mathrm{E_p log}\frac{\mathrm{p(x)}}{\mathrm{q(x)}} \tag{1.4}$$

.

The relative entropy is a measure of the distance between two probability distributions. In Statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy D (p||q) measures the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need H(p) + D (p||q) bits on the average to describe the random variable.

**Corresponding author:** *Parmil Kumar**
*Department of Statistics, University of Jammu, Jammu, India*

Nonetheless, it is often useful to think of relative entropy as a "distance" between distributions. Next we introduce the mutual information, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other random variable.Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information denoted by $I(X;\ Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product distributions p(x) p(y):

$$\mathrm{I(X;\ Y)} = \sum_{x\in X}\sum_{y\in Y} \mathrm{p(x,y)log_2}\frac{\mathrm{p(x,y)}}{\mathrm{p(x)q(y)}} \quad = \mathrm{D[p(x,y)||p(x)q(y)]}$$

$$= E_{p(x,y)} log \frac{p(x,y)}{p(x)q(y)} \tag{1.5}$$

## 2. TEXT CATEGORIZATION METHODS

Text categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data.

Text categorization may be formalized as the task of approximating the unknown target function $\emptyset : D \times C \rightarrow \{T, F\}$ (that describes how documents ought to be classified), by means of a function.

$\widehat{\emptyset}: D \times C \rightarrow \{T, F\}$ called the classifier, where $C = \{c_1, c_1, \ldots c_{|C|}\}$ is a predefined set of categories and $D$ is a (possibly infinite) set of documents. If $\emptyset(d_j, c_i) = T$, then $d_j$ is called a positive example (or a member) of $c_i$, while if $\emptyset(d_j, c_i) = F$, it is called a negative example of $c_i$. There are many methods employed for text categorization. Among them Feature selection has been extensively used in literature.

**Feature Selection Method**

Feature selection is an important step in TC (Yiming Yang, Jan O. Pedersen. 1997), in recent years a growing number of statistical classification methods and machine learning techniques have been applied for this task. The prevailing feature selection methods include document frequency (DF) thresholding, information gain (IG), and mutual information (MI). We will discuss DF in detail in following section.

Feature selection (also known as subset selection) is a process commonly used in machine learning, where in a subset of the features available from the data is selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to the accuracy; one discards the remaining, unimportant dimensions. This is an important stage of pre-processing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction).Two approaches namely Forward selection and Backward selection are pivotal in in Feature selection method.

**Forward selection approach**: It starts with no variables and adds the variables one by one. At each step of addition of the variable, the variable that decreases the error the most is added and process is repeated until any further addition does not significantly decrease the error.

**Backward selection approach:** It start with all the variables and eliminate them one by one, at each step of removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly. To reduce over fitting, the error referred to above is the error on a **validation set** that is distinct from the **training set**.

Other method that is also cited in literature is Document Frequency Thresholding.It is an another dimension of feature selection method.

**Document Frequency Thresholding**

Document frequency is the number of documents in which a term occurs. Only the terms that occur in a large number of documents are retained. Yang and Pedersen's experiments (1997) showed that it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness. DF thresholding is the simplest technique for vocabulary reduction. It scales easily to very large corpora with an approximately linear computational complexity in the number of training documents.

**Mutual Information**

Mutual information for feature selection in Text Categorization is defined as if a category *c and* a term *t* have probabilities $P(t)$ and $P(c)$ respectively, then their mutual information $I(t, c)$ is defined as:

$$I(t,c) = log_2 \frac{p(t,c)}{p(t)\times p(c)} = log_2 \frac{p(t \Lambda c)}{p(t)\times p(c)} \quad (2.1)$$

Here $I(t,c)$ compares the probability of observing t and c together with the probabilities of observing t and c independently. If there is a genuine association between t and c, then the joint probability P(t, c) will be much larger than $P(t) \times P(c)$, in this case $I(t,c) >> 0$. If there is no significance relationship between t and c, then $P(t,c) \approx P(t) \times P(c)$, and thus, $I(t,c) = 0$. If t and c are in complementary distribution, then $P(t,c)$ will be much less than $P(t) \times P(c)$, Forcing $I(t,c) << O$. According to Equation (3.1), the Mutual information of t and c can be negative, which is in conflict with the definition of Mutual information in information theory (Thomas & Cover, 1991) where it is always non-negative, so it would seem that the mutual information defined in (3.1) is not the one defined in information theory. WE will illustrate this with an example consider $p(t) = 0.9, p(c) = 0.6, P(t \wedge c) = 0.4$, then

$$I(t,c) = log_2 \frac{p(t \Lambda c)}{p(t) \times p(c)} = log_2 \frac{0.4}{0.9 \times 0.6} = log_2\, 0.74 < 0$$

While with usual definition as used in information theory $I(t,c) = -0.13$ .

In information theory (Thomas & Cover, 1991), the term "mutual information" refers to two random variables. It seems that the term “mutual information” has been used for something which should correctly be termed "pointwise mutual information" as it is applied not to two random variables (F. Sebastiani, 2002) but rather to two particular events from the sample spaces on which the two random variables are defined. This is the version used in current studies, and Equation (2.1) is really Pointwise Mutual Information (PMI). The "Mutual Information" method used for feature selection in Text Categorization should correctly be termed "Pointwise Mutual Information".

**Pointwise Mutual Information**

Mutual Information is a criterion commonly used in statistical language modeling of word associations and related applications (F. Sebastiani, 2002), (R. Fano, 1961). Many research workers have contributed a lot in understanding the applications of Mutual Information in Text Categorization (Chouchoulas and Q. Shen, 1999), (Yiming Yang, Xin Liu, 1999) as discussed above. It should correctly be termed "Pointwise Mutual Information" as it is not being applied to two random variables, since in information theory, the term "mutual information" refers to two random variables (Thomas & Cover, 1991).

Given a category c and a term t, let
A denote the number of times c and t occur together,
B denotes the number of times t occurs without c,
C denotes the number of times c occur without t, and N denotes the total number of documents in c.

The pointwise mutual information criterion between t and c is defined by () is as:

$$PI(t,c) = log_2 \frac{p(t \Lambda c)}{p(t)\times p(c)} \quad (2.2)$$

And this can be estimated using

$$PI(t,c) \approx log_2 \frac{A \times N}{(A+C)\times(A+B)} \quad (2.3)$$

These category-specific scores of a term are then combined to measure the goodness of the term at a global level.

Let $\{ci\}_{i=1}^{m}$ denote the set of categories in the target space. Typically it can be calculated in one of two ways
Either

$$PI_{avg}(t) = \sum_{i=1}^{m} p(c_i)\, I(t,c_i)\,, \quad (2.4)$$

Or

$$PI_{max}(t) = \max \{I(t,c_i)\},\ i = 1,2,3 \ldots\ldots m. \quad (2.5)$$

After the computation of these criteria, thresholding is performed to achieve the desired degree of feature elimination from the full vocabulary of a document corpus. According to (K. W. Church and P. Hanks, 1989), (R. Fano, 1961). in a general way, pointwise mutual information as defined () compares the probability of observing t and c together (the joint probability) with the probabilities of observing t and c independently (chance). If there is a genuine association between t and c, then the joint probability P(t,c) will be much larger than chance $P(t)P(c)$, and consequently $PI(t,c) >> 0$. If there is no significant relationship between t and c, then $P(t,c) \approx P(t)P(c)$, and thus, $PI(t,c) \approx 0$.

If t and c are in complementary distribution, then $P(t,c)$ will be much less than $P(t)P(c)$, forcing $Pl(t,c) << O$. That is, pointwise mutual information as defined above can be negative, $PI(t)$ avg also can be negative, in (reference), $PI(t)$ avg is found be negative for about 20% of the terms.

According to information theory, the MI of any random variables X and Y is always non-negative, so the pointwise mutual information as defined above is not actually the "mutual information" as defined in information theory. Next we will discuss the concept of MI in information theory (Thomas & Cover).

**Information Theoretic Mutual Information**
The MI between two discrete random variables X and Y is defined as

$$\text{I (X; Y)} = \sum_{x\in X}\sum_{y\in Y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} \quad (2.6)$$

Where $p(x,y)$ and $p(x)$ , $p(y)$ are the joint and the marginal probability distribution.

This measure satisfy the properties of symmetry i.e

$$\text{I (X; Y)} = \text{I (Y; X)} \quad (2.7)$$

$$\text{I (X; Y)} = \text{H(X)} + \text{H(Y)} - \text{H(X, Y)} \quad (2.8)$$

And (Non-negativity of mutual information): $\text{I (X; Y)} \geq 0$ with equality if and only if *X* and *Y* are independent.

Where $H(X)$ is the entropy of the random variable X. *H*(*Y*) is the entropy of the random variable Y and $H(X;\ Y)$ is the joint entropy of these variables. The category-specific term scores are combined to measure the goodness of the term *t* at a global level.

Let $\{C_i\}_{i=1}^m$ denote the set of categories in the target space,$= \cup_{i=1}^m c_i$ , *let* $T = \{t, \bar{t}\}$ denote the set in which term *t* occurs or *t* does not occur. Using equation (), the mutual information between T and C is defined as:

$$I(T;C) = \sum_{i\in T}\sum_{c_i\in C} P(t', c_i) log_2 \frac{P(t', c_i)}{P(t')P(c_i)} \quad (2.9)$$

$$= \sum_{i=1}^m P(t\Lambda c_i) log_2 \frac{P(t\Lambda c_i)}{P(t)P(c_i)} + \sum_{i=1}^m P(\bar{t}\Lambda c_i) log_2 \frac{P(\bar{t}\Lambda c_i)}{P(\bar{t})P(c_i)} \quad (2.10)$$

I (X; Y) is the MI criterion between T and C,

As MI is non-negative i.e. $\text{I (X; Y)} \geq 0$

Let $\text{I (t)} = \text{I (T; C)}$

$$I(t) = \sum_{i=1}^m P(t\Lambda c_i) log_2 \frac{P(t\Lambda c_i)}{P(t)\times P(c_i)} + \sum_{i=1}^m P(\bar{t}\Lambda c_i) log_2 \frac{P(\bar{t}\Lambda c_i)}{P(\bar{t})\times P(c_i)} \quad (2.11)$$

$So\ \ I\ (t) \geq 0$ .

On the line of Information Theoretic Mutual Information,

Given a training corpus, for each unique term *t* one can compute the MI and then remove from the feature space those terms whose information gain is less than some predetermined threshold, this is the MI method. Here MI compares the probability of observing t and c together with the probabilities of observing t and c independently.

**3. PERFORMANCE OF KNN CLASSIFIER**
The technique of feature selection is commonly employed by online search engine, spam mail filtration and data mining. With the tremendous increase in information available in digital information these technique has become handy like collecting the diamonds from the mine. Information scientists, Communication Engineer, Knowledge management worker are contributing a lot along with mathematician and statistician on the development of this area of Information theory. Various authors have done significant work in this direction viz (Y. Liu, 2004) (St. M. Yang, X.-B. Wu, Z.-H. Deng, M. Zhang and D.-Q. Yang, 2002). Many existing experiments show IG is one of the most effective methods (Y. Yang, J. and O. Pedersen. 1997), by contrast and MI has been demonstrated to have relatively poor performance than others but in this paper we show that the performance of corrected MI method is similar to that of Information Gain and it is considerably better than PMI.

**Information Gain**

Information Gain is the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. Let m $\{C_i\}_{i=1}^{m}$ denote the set of categories in the target space. The information gain of term t is defined to be:

$$G(t) = -\sum_{i=i}^{m} p_r(c_i) \log p_r(c_i) + p_r(t) \sum_{i=1}^{m} p_r(c_i|t) \log p_r(c_i|t) + p_r(\bar{t}) \sum_{i=1}^{m} p_r(c_i|\bar{t}) \log p_r(c_i|\bar{t}) \quad (3.1)$$

Given a training corpus, the information gain is computed for each unique term. Those terms whose information gain is less than some predetermined threshold are removed from the feature space.

**kNN Classifier**

Once the Information is available in digital form the need arises to category it in the desired classes so that junk or irrelevant results can be ignored. Among the most commonly used methods of classification viz. Naïve bayes, K-nearest neighbors, Decision trees, Rocchio's algorithm, Support vector machines, Neural networks, Linear least squares fit. kNN Algorithm is commonly used Text Categorization [F. Sebastiani, 2002

In order to compare the modified feature selection methods to the original ones, we employed the k-nearest neighbor classifier (kNN), which is commonly used in the text categorization field. Given an input document represented as sparse vector of word weights, the classifier comes up with a list of confidence scores for all categories and assigns the input to the category with the highest confidence score.

To generate the list, a kNN classifier first determines the k nearest neighbor to the input among all the training documents, where similarity of each documents to the input is measured by the cosine between the two documents vectors. These k similarity weights are then summed by the category to form the confidence list.

We choose kNN because it is extremely suitable for our experiments. Firstly, it's one of the top-performing classifiers. Evaluations have shown that it outperforms nearly all the others systems; except for the SVM system, which is not suitable for the test because its result will not be affected by the feature selection process. Secondly, because it scales well to a large number of features, the kNN system can be used to examine all degree of feature selection, observe the effects on accuracy. Finally, the kNN classifier is context sensitive, thus enabling a better observation on feature selection process.

**K-Nearest Neighbors**

The ***k*-nearest neighbor algorithm** is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors, where *k* is a positive integer, typically small. If *k* = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose *k* to be an odd number to avoid tied classes.

The purpose of kNN algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and are only based on the memory. Foe a given query point, we find K number of objects or (training points) closest to the query point. The classification done by is using majority vote among the classification of the K objects. Any ties can be broken at random. K Nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance.

The kNN classifier is also used to classify whether a data point is normal or abnormal. kNN classifier has been first used in instrusion detection area for anomaly detection to learn programme behavior and uncover instrusion from audit data.

Similar to Mahalanobis distance formula the kNN classifier measure the distance between two data points P and Q by Euclidian distance. The distance actually represented their degree of similarity. The shortest the distance between them, the maximum similar they are. Mathematically speaking,

$$(P, Q) = \sqrt{\sum_{i=0}^{N} (p_i - q_i)^2} \quad (3.2)$$

Where $p_i$ and $q_i$ are the values of the $i^{th}$ attribute of the point p and q respectively.

The data for KNN algorithm consist of several multivariate attributes name $X_i$ that will be used to classify Y. The data of KNN can be on any measurement scale from ordinal, nominal, to quantitative scale but here we will deal with only quantitative $X_i$ and binary (nominal) Y.

Step by step method on how to compute K-nearest neighbors using kNN algorithm:

1) Determine the parameter K = number of nearest neighbors.
2) Calculate the distance between the query-instance and all the training samples.
3) Sort or Rank the distance and determine the nearest neighbors based on the K-th minimum distance.
4) Gather the category of the nearest neighbors.
5) Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

## 4. RESULT AND CONCLUSION

Next, we illustrate the above along with the help of an example. We have secondary data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here are four training samples:

| X1 = Acid Durability (seconds) | X2=Strength(kg/square meter) | Y = Classification |
|---|---|---|
| 7 | 7 | Good |
| 3 | 4 | Good |
| 1 | 3 | Bad |
| 7 | 4 | Bad |

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can it be guess of what the classification of this new tissue is?

1) Determine the parameter K = number of nearest neighbors Suppose use K = 3

2) Calculate the distance between the query-instance and all the training samples Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

3) Sort and Rank the distance and determine nearest neighbors based on the K-th minimum distance.

| X1 = Acid Durability (seconds) | X2 = Strength (kg/square meter) | **Square Distance to query instance (3, 7)** |
|---|---|---|
| 7 | 7 | $(7-3)^2+(7-7)^2=16$ |
| 3 | 4 | $(3-3)^2+(4-7)^2=9$ |
| 1 | 3 | $(1-3)^2+(3-7)^2=20$ |
| 7 | 4 | $(7-3)^2+(4-7)^2=25$ |

| X1=Acid Durability (seconds) | X2 = Strength (kg/square meter) | Square Distance to query instance (3, 7) | **Rank minimum distance** | **Is it included in 3-Nearest neighbors?** |
|---|---|---|---|---|
| 7 | 7 | $(7-3)^2+(7-7)^2=16$ | 2 | Yes |
| 3 | 4 | $(3-3)^2+(4-7)^2=9$ | 1 | Yes |
| 1 | 3 | $(1-3)^2+(3-7)^2=20$ | 3 | Yes |
| 7 | 4 | $(7-3)^2+(4-7)^2=25$ | 4 | No |

4. Gather the category of the nearest neighbors. Notice in the second row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

| X1=Acid Durability(seconds) | X2=Strength (kg/square meter) | Square Distance to query instance (3, 7) | **Rank minimum distance** | **Is it included in 3-Nearest neighbors?** | **Y = Category of nearest Neighbor** |
|---|---|---|---|---|---|
| 7 | 7 | $(7-3)^2+(7-7)^2=16$ | 2 | Yes | Good |
| 3 | 4 | $(3-3)^2+(4-7)^2=9$ | 1 | Yes | Good |
| 1 | 3 | $(1-3)^2+(3-7)^2=20$ | 3 | Yes | Bad |
| 7 | 4 | $(7-3)^2+(4-7)^2=25$ | 4 | No | - |

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

Here we have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in Good category.

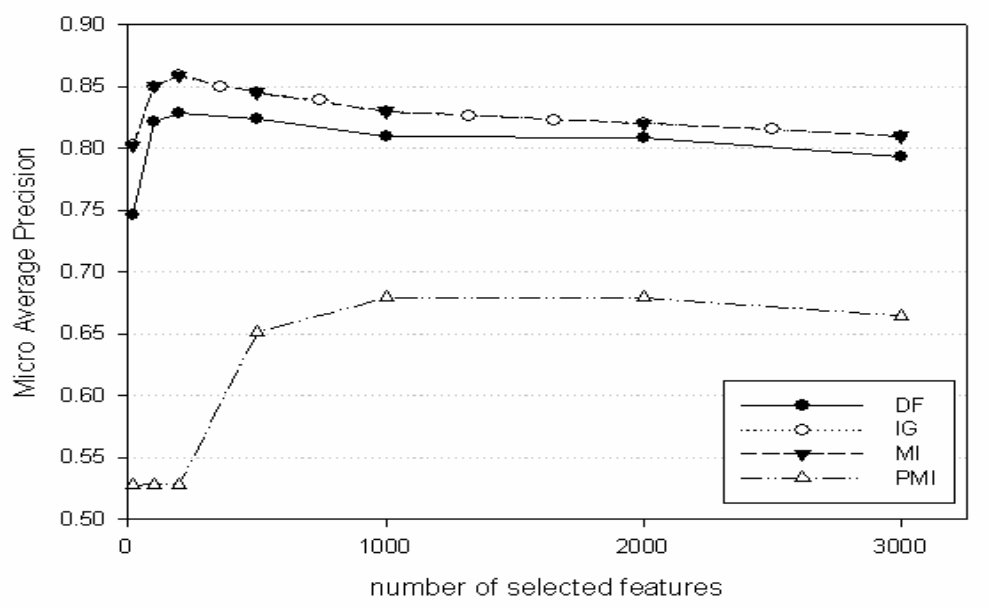
Next, we have shown that PMI has lowest performance when compared with other feature selection method.But information Theoretic Mutual Information selection method outperform the all selection methods.

**Data Collections**

A corpus used in our experiments is Reuters-21578 collection [http://www.daviddlewis.com/resources/testcollections/reuters21578/].

The Reuters-21578 collection is the original Reuters-22173 with 595 documents which are exact duplicates removed, and has become a new benchmark lately in text categorization evaluations. In our experiment, we only consider those documents that had just one topic, and the topics that have at least 5 documents. The training set has 5273 documents, the testing set has 1767 documents.

Below Figure show the performance curve of kNN on the Reuters-21578 collection after feature selection using DF, IG, PMI, and MI. It can be seen in Figure that the MI method outperforms the PMI method.




We have compared the original DF, IG, MI and PMI. A number of statistical classification and machine learning techniques have been applied to text categorization, along with kNN, which is one of the top-performing classifiers, and evaluations have shown that it outperforms nearly all the other systems.We find that IG and MI are the most effective in our experiments, that is, IG and MI produce similar performance of the classifiers. DF thresholding performed almost similar. In contrast, PMI has by far the lowest performance.

**REFERENCES**

1. C. E. Shannon (1948).”A Mathematical Theory of Communication”, *Bell System Tech. Journal, vol. 27,* pp. 379-423.
2. Chouchoulas and Q. Shen (1999). A Rough Set-Based Approach to Text Classification. *Proceedings of the 7th International Workshop on Rough Sets*, pages 118-127
3. Dai Liu-ling, Huang He-yan, Chen Zhao-Xiong (2005). A comparative Study on Feature Selection in Chinese Text Categorization, *Journal of Chinese Information Processing. Vol.18* No.1:26-32.
4. E. Wiener, J.O. Pedersen, and A. S. Weigend. (1995). A Neural Network Approach to Topic Spotting. http://www.cs.usyd.edu.au/~comp4302/liz.pdf
5. F. Sebastian (2002), Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34(1)*:1-47..
6. H. Zhang (2004).The optimality of naive Bayes. *The 17th International FLAIRS conference, Miami Beach*, May 17-19, 2004.
7. J.R. Quinlan (1986). Induction of Decision Trees. *Machine Learning, 1(1)*: pp.81-106.
8. Joachims, T., (1998). Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98*, 10th European Conference on Machine Learning, eds. C. N´edellec & C. Rouveirol, Springer Verlag, Heidelberg, DE: Chemnitz, DE, pp. 137–142,. Published in the “Lecture Notes in Computer Science” series, number 1398.
9. K. W. Church and P. Hanks.(1990) Word Association Norms, Mutual Information and Lexicography,Computational Linguistics,Vol 16(1),22-39.
10. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A. 1999. Rough sets: A tutorial. A New Trend in Decision-Making, *Springer-Verlag, Singapore*, 3-98

11. Pawlak Z. (1982) Rough Sets. *International Journal of Computer and Information Science,, 11(5):* 341-356
12. R. Fano. (1961). Transmission of Information. *MIT Press, Cambridge, MA*
13. Reuters21578. [http://www.daviddlewis.com/resources/testcollections/reuters21578/].
14. S. Doan and S. Horiguchi, (2005) "An Efficient Feature Selection using Multi- Criteria in Text Categorization for Naïve. Proceedings of the 4th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering Data Bases Salzburg, Austria.
15. Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Inform. Process. Man. 24, 5, 513–523.
16. Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys, 34(1),* pp. 1–47,
17. Shan Songwei, Feng Shicong, Li Xiaoming. (2003). A Comparative Study on Several Typical Feature Selection Methods for Chinese Web Page Categorization, *Journal of the Computer Engineering and Application, Vol.39* No.22 :146-148.
18. Stewart M.Yang, Xiao-Bin Wu, Zhi-Hong Deng, Ming Zhang, Dong-Qing Yang (2002). Modification of Feature Selection Methods Using Relative Term Frequency. *Proceedings of ICMLC-2002,* pp. 1432-1436.
19. T. M. Cover and J. A. Thomas, (1996) Elements of Information Theory, *John Wiley & Sons, Inc.* Print ISBN.
20. T. Mitchell.(1997) Machine Learning. McCraw Hill, New York.
21. Y. Liu,(2004) A Comparative Study on Feature Selection Methods for Drug Discovery, J. Chem. Inf. Comput. Sci.
22. Y. Yang and X. Liu.( 1999) A re-examination of text categorization methods. *(SIGIR'99),* pp. 42-49.
23. Yiming Yang, Jan O. Pedersen. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of ICML-97,* pp. 412-420.
24. Yiming Yang (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, Vol. 1,* No. 1/2, pp 67–88.


**Source of support: Nil, Conflict of interest: None Declared**