ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Enriching Text Classification Using Fuzzy and Probability Function

Ganesh Khot [1], Namrata Mahajan [2], Swapna Bhise[3], Gargi Joshi[4]

P.G. Student, Department of Information Technology Engineering, DYPCOE, Pune, Maharashtra, India[1]

P.G. Student, Department of Information Technology Engineering, DYPCOE, Pune, Maharashtra, India[2]

P.G. Student, Department of Information Technology Engineering, DYPCOE, Pune, Maharashtra, India[3]

Associate Professor, Department of Information Technology Engineering, DYPCOE, Pune, Maharashtra, India[4]

ABSTRACT: Text Classification is very important research field in information fetching and text mining. Text classification approach is gaining more importance because of the receptiveness of large number of e-documents from a variety of resource. The main task in text categorization is to assign text documents in predefined categories based on documents contents. Since word detection is a difficult and time consuming task Bayesian text classifier is an appropriate method to deal with different word forms and new words. Also, fuzzy theory may be used to manage ambiguity. Semi-supervised learning is the main approach to the problem. In early days labelled data (label pairs) was used by some classifiers. Labelled data are very hard to gain while unlabeled data is easy and smooth, therefore semi-supervised learning is a good idea to demote human labour and increase accuracy. Labelled data or labelling cost seems to be high. Therefore, the proposed system is being planned to work on by using unlabeled data. The proposed system plans to a semi-supervised learning with universum learning based approaches. Universum, is a collection set of non-examples that do not belong to any class of interest. If text classification does not occur then the huge and large data cannot be classified. The most of the existing systems are in the market to do are Ada- boost technique, Naive bayes, support vector machine, neural , Particle Swarm Optimization (PSO), etc.. Fuzzy ANN and Bayesian probability algorithm is the process of semi-supervised text categorization and results obtained are encouraging. The experimental results are obtained by Shannon- info gain , Tf-idf, K-means , Gaussian distribution , Fuzzy ANN , Bayesian probability , Atkinson index methodologies. The results output to an text classification rule and text clusters .

KEYWORDS: Fuzzy ANN, Probability functions, learning with universum, text classification.

I. INTRODUCTION

Text classification is one of the unique quality technique in text mining to categorize the documents in a supervised form. The procedure of text classification involves two main problems are the Fuzzy Artificial neural network and probability function measures in the training sequence and then the actual classification of the document using these feature terms in the test sequence. Semi-supervised learning is the main method to this problem. However, problem in applying supervised learning methods to real-world problems is the cost of obtaining sufficient unlabelled training data, since supervised learning methods often require a huge, precludive, number of unlabeled training examples to learn accurately. Labelling is time-consuming ,costly and typically it is done manually. Conversely, unlabelled data is somewhat easy to collect, and many algorithms and experimental results have demonstrated that it can considerably increase learning accuracy in specific practical problems . Consequently, semi-supervised learning, which involves learning from a set of elements of both labelled and unlabelled data, has recently attracted significant interest .

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

II. RELATED WORK

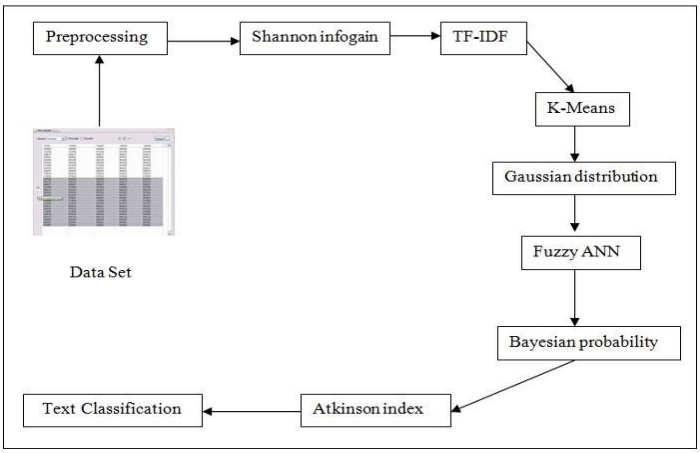

Fig: System Model

Pre-processing

Pre-processing is conducted, where string is processed to its basic meaning words by the 4 main activities: Sentence Segmentation , Tokenization , Removing Stop Words and Word Stemming.

- Sentence Segmentation is location detection and separating source text into sentence.
- Tokenization is separating the input problem into individual or single words.
- Stop word removal in any document narration the conjunction words does not play much role in the meaning of the document, so by removing these words (like: is, the, for, an) from the documents which greatly deducts the overhead of processing
- Stemming many of the elongated words in the English language generally fail to provide exact meaning in the given scene and also they increases the computational time. So it is necessary to bring the words to their base form by replacing its extended characters with desired characters (Example: studied to study, where ied is replaced with y). And this can be represent with the following algorithm

ALGORITHM 1: PREPROCESSING

Step 0: Start
Step 1: Identify and read string
Step 2: Divide string into records on space and store in a vector V
Step 3: Remove Special Symbols
Step 4: Identify Stop words
Step 5: Remove Stop words
Step 6: Identify Stemming Substring
Step 7: Replace Substring to desire String
Step 8: Concatenate Strings
Step 9: stop

Shannon Info-Gain

Shannon info-gain is the process to summarize each of documents in an Information Ratio result, we use Shannon's term weighting based on Information Gain Ratio (IGR).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

This method decoctes the similarity structure among a set of documents through a hierarchical clustering, then gives higher weights to words . Important words are calculated based on IGR as follows

$$IGR(C) = -\sum (|C_i| / |C|) \log (|C_i| / |C|) \quad ....(2)$$

Where $C_i$is the frequency of the word *w* in Cluster C.

TF-IDF

This step is next to Shannon info-gain process, now here the documents are been clustered using simple centroid techniques using the content of the documents itself. For this we can use TF-IDF. The most repeated word in the document actually plays a vital role to recognize the importance of the sentence. Then rank of the sentences can be computed as the sum of rank of all words in that sentence. The rank of any words w_i can be finalized by the calculation of tf-idfs as shown below in equation .

$$W_i = tf \; x \; idfi = tf_i x \; log \; (N / ni)$$

Where tf_i is the TF of word *i* in the document, N is the total number of documents, and n_i is number of documents in which word i occurs.

Gaussian Distribution

This step actually adds the more amount of words into the clusters based on the Gaussian distribution which is presented by Gaussian kernel equation.

Gaussian Kernel Equation

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \text{----------(2)}$$

μ= mean of distribution
σ^2 = variance of distribution
y= continuous variable
P(y)= probability of y

Finding the global minimum (or maximum) of a function is far more difficult: symbolic (analytical) methods are oftenly not applicable and the use of numerical solution strategies often leads to very hard challenges. This can be conviniently solve by using Gaussian distribution model as mentioned in the below algorithm.

Step 1: Generate N solutions x_i(i = 1, 2, … ,n) from $[\alpha_i, b_i]$, using the uniform design technique.

Step 2: Assess the fitness of all individuals in the initial population and retain the best solution.

Step 3: Order the population by fitness in decreasing order sorting and choose the optimal m individuals (m ≤ N).

Step 4: Analyze the generated m individuals information and calculate the mean u_i^ and variance $\sigma_{i^\wedge}$ of each variable.

Step 5: Sample N new solutions according to formula (2).

Step 6: If the given end condition (up to the required number of iterations n_{max}) is not met, go to step 2.

K-Means Clustering

K-means clustering is one of the unsupervised reckoning methods used to group similar objects in to smaller portions called clusters so that similar objects are gathered together . The algorithm aims to minimize the under hold within cluster variance and maximize the intra cluster's variance. The technique involves determining the number of clusters at first and randomly designating cluster centroids to each cluster from the whole datasets; this step is called initial part of cluster centroids. The distance between each point in the whole datasets and every cluster centroid is then computed using a distance metric (e.g. Euclidean distance) . Then, for every data point, the minimum distance is identified and that point is designated to the closest cluster. This step is called cluster assignment, and is iterated until all of the data

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

points have been assigned to one of the clusters. Finally, the mean for each cluster is calculated based on the gathered values of points in each cluster and the number of points in that cluster. Those mean's are then designated as new cluster centroids, and the process of finding distances between each point and the new centroids is iterated, where points are re-assigned to the new closest clusters. The process repeats for a fixed number of times, or until points in each cluster stop moving across to different clusters. This is called convergence. The steps of the K-means can be summarized in Figure below.

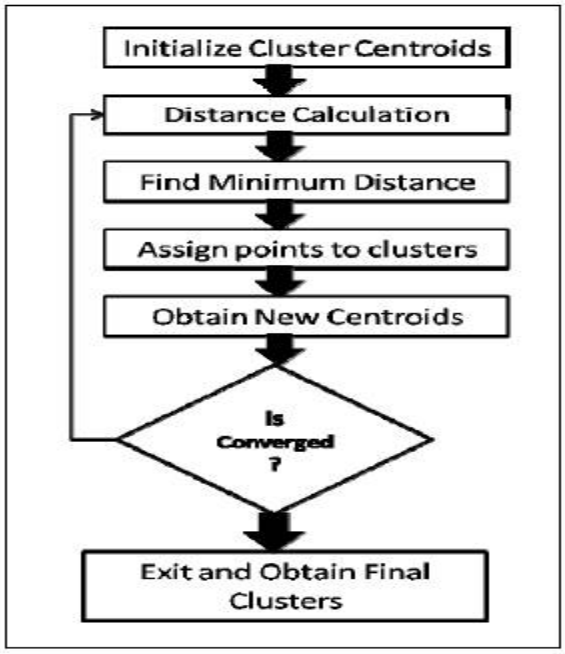

Fig. K means Computational Steps

Fuzzy ANN

ANN:Artificial intelligence systems are working on the basis of biological neurons in the brain. So biological neurons are come to a conclusion by inter communicating with the many neurons. In the same way in our proposed theory mean and standard deviations are calculated for the attributes of the clusters which are formed using K-Means in the previous step.

$$f(\mu) = \frac{\sum_{n=1}^{\infty}(A_i)}{n} \quad \text{----------1}$$

$$f(\delta) = \sqrt{\frac{\sum_{n=1}^{\infty}(\mu - A_i)}{n}} \quad \text{----------2}$$

Then by using this mean and standard deviation our system identifies the ranges based on the equation 3 and 4.

Min range = $f(\mu) - f(\delta)$ ---------3

Max range = $f(\mu) + f(\delta)$---------4

Then new and hidden neuron clusters will be formed based on the ranges of equation 3 and 4. These new clusters are more crisp and having perfect sharp values in the given ranges.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

The working pattern of ANN can be depicted in the figure no 2

.

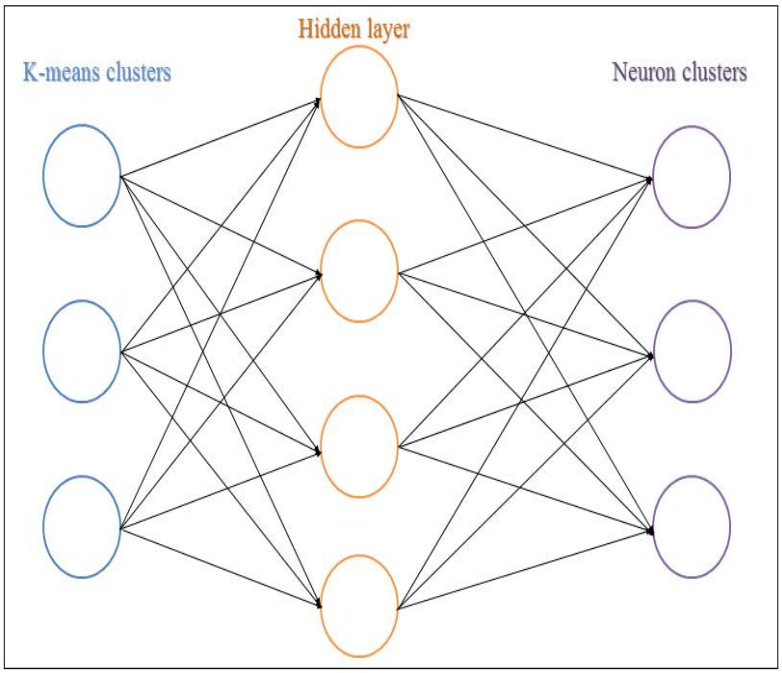

Figure 2: Working model of ANN

Fuzzy Logic : It is a form of abstract classification theory which works on the basis of crisp values like Very low, low, medium, high and very high. Fine clusters which are formed by the ANN now feed to the fuzzy logic for the classification of the data for the DOS attack. Based on the crisp values and the cluster weights data for different attacks been classified in abstract level.

Bayesian Probability

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

There are two types of probabilities –

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

$P(H/X)= P(X/H)P(H) / P(X)$

Algorithm can be represented as shown below

1. Choose the set of relevant variables $\{X_i\}$ that describe the domain.
2. Choose an ordering for the variables, $< X_1……..X_n>$.
3. While there are variables left.

(a) Add the next variable Xi to the network.

(b) Add arcs to the Xi node from some minimal set of nodes already in the net,

Parents(X_i), such that the following conditional independence property is satisfied:

$P(X_i \mid X'_1,,,,,,,,,,,,X'm) = P(X_i \mid Parents(X_i))$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

where X'_1 ,,,,,,,,,,,,,,X'_m are all the variables preceding X_i.
(c) Define the probability for Xi.

Bayesian Belief Network: Bayesian Belief Networks specify joint conditional probability distributions. They are also called as Belief Networks, Bayesian Networks, or Probabilistic Networks.
There are two components that define a Bayesian Belief Network –

- Directed acyclic graph
- A set of conditional probability tables
- Atkinson Index

Here in this step data swapping is originated from the fact of unequal data distribution among the clusters for proper classification. For which system uses the Atkinson indices as the effective mode of distribution identification as indicated in equation 1.

(1) Atkinson Index calculation can given by

$$A_\varepsilon(y_1,\ldots,y_N) = \begin{cases} 1 - \frac{1}{\mu}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^{1-\varepsilon}\right)^{1/(1-\varepsilon)} & \text{for } 0 \le \epsilon \ne 1 \\ 1 - \frac{1}{\mu}\left(\prod_{i=1}^{N} y_i\right)^{1/N} & \text{for } \varepsilon = 1, \end{cases}$$

Where y_i is individual cluster load (i = 1, 2, ..., N) and μ is the mean load

III. PROPOSED SYSTEM

The proposed system takes input an Reuters dataset an text document for classification rule output. The input document is pre-processed using stop words, stemming and tokenization techniques .Then the second step where semi-supervised algorithms like Shannon-info gain gain's required important data from the given output text document. The obtained information is then computed for its term frequency and inverse document frequency i.e TF-IDF. The calculated term frequencies are clustered and distributed by Gaussian technique. Then using atkinson index the distributed data is indexed .The formed index is further implied as an input for Fuzzy ANN algorithm to get the required output of classified text.

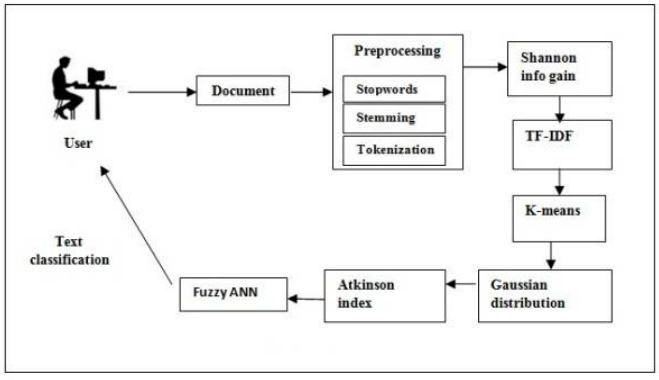

Fig:System Architecture

PROCESS SUMMARY

1] Unlabelled Reuters data is used as input data by user. The input data is in XML format with the type text.
2] The input Reuters data is pre-processed for sorting out important data. Sorting is done by using stop words, stemming and tokenization techniques.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

3] Shannon-info gain algorithm is used to gain important data within the input document after the sorting pre-processing technique.
4] The gained data within the document is calculated for its term frequency and inverse document frequency.
5] Calculated frequencies are then clustered by using k-means clustering algorithm.
6] The clustered data is then distributed by Gaussian distribution technique.
7] Atkinson index approach is performed to index the distributed data of the input data.
8] Last but not the least Fuzzy ANN algorithm is implemented on the index formed in the previous process for the desired output of classified text.

IV. EXPERIMENTAL RESULTS

Some experimental evaluations are performed to show the effectiveness of the system. And these experiments are conducted on windows based java machine with universally used IDE Net-beans. Also the numbers of matched class labels for classified data form the data set is used to set benchmark for performance evaluation.
Numbers of relevant identified class labels from the dataset is used to show the effectiveness of the system.
Below are the definition of the used measuring techniques i.e. precision and recall.

Precision: it is a ratio of numbers of proper identified class labels to the sum of total numbers of relevant and irrelevant identified class labels. Relative effectiveness of the system is well expressed by using precision parameters.

Recall: it is a ratio of total numbers of relevant identified class labels retrieved to the total numbers of relevant class labels not retrieved. Absolute accuracy of the system is well narrated by using recall parameter
Numbers of scenarios presents where one measuring parameter dominates the other. by taking such parameters into consideration we used two measuring parameters such as precision and recall.
For more clarity let we assign

- X = the number of relevant class labels retrieved,
- Y = the number of relevant class labels are not retrieved, and
- Z = The number of irrelevant class labels retrieved .

So, Precision = (X/ (X+ Z))*100
And Recall = (X/ (X+ Y))*100

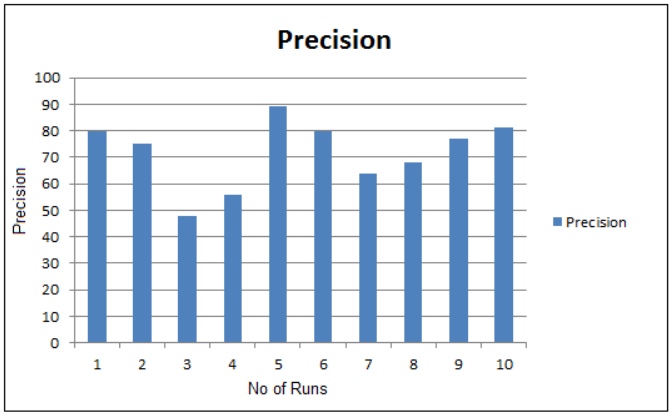

Fig.3. Average precision of the proposed Text classification detection method

In Fig. 3, by observing figure 3 it is clear that the average precision obtained by using text classification detection method is approximately 71.8%.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

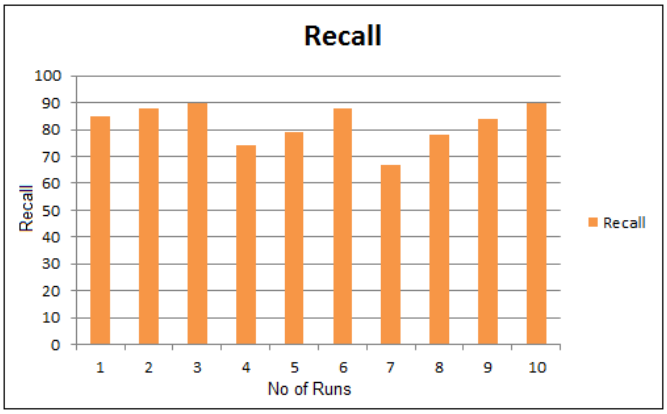

Fig.4. Average Recall of the proposed Text classification detection method

In Fig. 4, figure shows that the system gives 82.3% recall for the text classification method. By comparing these two graphs we can conclude that the text classification method gives high recall value compare to the precision value.

COMPARATIVE ANALYSIS WITH EXISTING SYSTEM

Existing system (Ada-boost)	**Proposed System (Fuzzy & Bayesian)**
Classification rule is not formed.	**Classification rule is formed**
Only unlabelled data is used.	**Labelled (random) type of data is used.**
Do not deal with class imbalance in minority for misclassified data	**It helpfully deals with class imbalance in minority for misclassified data.**

V. CONCLUSION

The proposed system concludes that the Ada-Boost approach which results to training error and normalization factor and not classification rule. Also the technique results to classification loss on target examples.[1] The system used is fuzzy L-R type number which only can categorize Persian language text document. Also evaluates precision recall parameters.[2][3].Our system tries to surmount all these factors by using probability techniques. The experimental results to best performance, time efficient and accuracy. In future a time ahead ,we will try to work on the domain of image classification and can input the other document extensions like doc.,pdf.,txt. Data.

REFERENCES

[1]Semi-Supervised Text Classification With Universum Learning. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Tao-Hsing Chang, and Tsung-Hsun Kuo. 2168-2267 _c 2015 IEEE.
http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
[2] Using Fuzzy LR Numbers in Bayesian Text Classifier for Classifying Persian Text Documents. Parisa Pourhassan Parisa_pourhassan@yahoo.com
International Journal of Information, Security and System Management, 2013, Vol.2, No.1, pp. 118-123
[3] A Fuzzy Based Approach for Multilabel Text Categorization and Similar Document Retrieval,Volume 5, Issue 9, September 2015 ISSN: 2277 128X
[4]W. Hu, J. Gao, Y. Wang, O. Wu, and S. J. Maybank, "Online AdaBoostbased parameterized methods for dynamic distributed network intrusion detection," IEEE Trans. Cybern., vol. 44, no. 1, pp. 66–82, Jan. 2014.[Online]. Available: http://dx.doi.org/10.1109/TCYB.2013.2247592.
[5]K. Zhou, X. Gui-Rong, Q. Yang, and Y. Yu, "Learning with positive and unlabeled examples using topic-sensitive PLSA," IEEE Trans. Knowl.Data Eng., vol. 22, no. 1, pp. 46–58, Jan. 2010.
[6]M. Sokolova and L. Guy, "A systematic analysis of performance measures for catagorization tasks," Inf. Process. Manage., vol. 45,pp. 427–437, Jul. 2009.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

[7]X. Shi, B. L. Tseng, and L. A. Adamic, "Information diffusion in computer science citation networks," in Proc. Int. Conf. Weblogs Soc. Media(ICWSM), San Jose, CA, USA, 2009, pp. 319–322.
[8]J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the Universum," in Proc. 23rd Int. Conf. Mach Learn. (ICML),Pittsburgh, PA, USA, 2006, pp. 1009–1016.
[9]W. Hu, W. Hu, and S. J. Maybank, "AdaBoost-based algorithm for network intrusion detection," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol.38, no. 2, pp. 577–583, Apr. 2008.X. Carreras, L. Márquez, and L. Padró, "Named entity extraction usingAdaBoost," in Proc. 6th Conf. Nat Lang. Learn. (COLING-02), Taipei,Taiwan, 2002, pp. 1–4.