ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

A Survey on Enriching Text Classification Using Fuzzy and Probability Function

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

BE. Student, Department of IT, DYPCOE, Ambi, Pune, India[1]

BE. Student, Department of IT, DYPCOE, Ambi, Pune, India[2]

BE. Student, Department of IT, DYPCOE, Ambi, Pune, India[3]

Asst. Professor, Department of IT, DYPCOE, Ambi, Pune, India[4]

ABSTRACT: Semi-supervised learning is a rare quality of classification process . Traditional classifiers use labeled data (label pairs). Labeled data are hard to obtain while unlabeled data is huge, therefore semi-supervised learning is a good idea to reduce human labour and improve accuracy. Labeled data or labelling cost seems to be high. Therefore, the proposed system is being designed by using unlabeled data. The proposed system design to a semi-supervised learning with universum learning based techniques. Universum, is a set of non-examples that do not belong to any class of interest. If text classification does not exist then the huge 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 Swaram Optimization (PSO), etc.. Fuzzy Ann and Bayesian probalility algorithm is the procedure of semi-supervised text catagorization and results obtained are encouraging. The experimental results are obtained by Shanon- infogain , Tf-idf, K-means , Gaussian distribution , Fuzzy ANN , Baysian probability , Atkinson index methodologies. The key points favourable to success for the proposed system are a) Improvement in performance of any semi-supervised laearning algorithm with multitude of unlabeled data b) efficient calculation by repetitive of boosting technique and exploiting both manifold and cluster assumption in training classification models.

General terms Semi-supervised learning algorithms, machine learning, artificial neural networks.

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

I. INTRODUCTION

text classification is one of the key technique in text mining to categorize the documents in a supervised manner. the procedure of text classification involves two main problems are the fuzzy artificial neural network and probability function measures in the training phase and then the actual classification of the document using these feature terms in the test phase. semi-supervised learning is the main approach 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. labeling is time-consuming ,costly and typically it is done manually. conversely, unlabelled data is relatively easy to collect, and many algorithms and experimental results have demonstrated that it can considerably improve learning accuracy in certain practical problems . consequently, semi-supervised learning, which involves learning from a combination of both unlabelled and unlabelled data, has recently attracted significant interest.

DOI: 10.15680/IJIRCCE.2016. 0402197

II. RELATED WORK

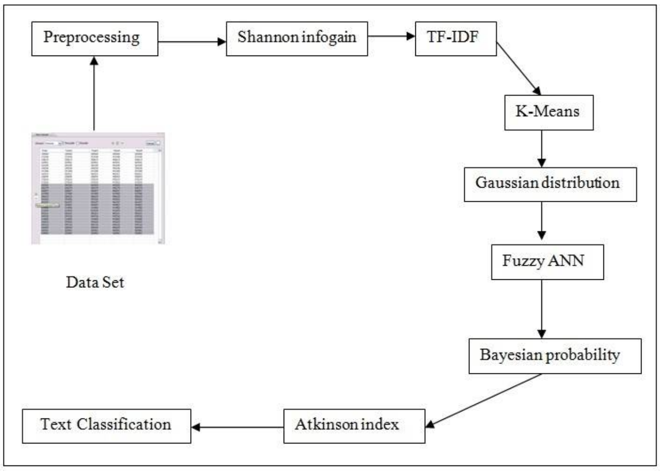

Fig :System Overview

Shannon info-gain algorithm

ID3 functions on information gain as its quality selection measure. This measure is based on pioneering work by Claude Shannon on information theory. Let node N holds D. The attribute with highest information gain is chosen as splitting attribute for node N. This attribute reduces the information needed to classify the tuples which results to the least randomness in these partitions. The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

where pi is the non-zero probability that an arbitrary tuple in D belongs to class Ci and is estimated by . A log function to the base 2 is used, because the information is encoded in bits. Info (D) is just the average amount of information needed to identify the class label of a tuple in D. Info (D) is also known as the entropy of D.

How much more information would we still need (after the partitioning) to arrive at an exact classification? This amount is measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term acts as the weight of the j th partition which is the expected information required to classify a tuple from D based on the partitioning by A. Information gain is defined as the difference between the original information

required (i.e., based on just proportion and the new requirement(i.e. obtained after partitioning on A).

$$Gain(A) = Info(D) - Info_A(D).$$

That is In other words, Gain(A) shows us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.

TF-IDF Algorithm

TF-IDF is the most common method used to describe documents in the Vector Space Model. It is composed of Term Frequency and Inverse Document Frequency. Term frequency is a lexical or dynamic weight of a term on a document achieved by counting the number of times particular term occurs in a text document. Inverse Document Frequency (IDF) is the count of all documents in the specific demography is divided by the number of documents that contains at least a single occurrence of the query term. The IDF gives a worldwide view of the term across the entire corpus, the lower the IDF value the more significant the term will be and it is calculate during (1).To get an effective term weight score,

$$idf = \log\left(\frac{N}{df_i}\right) \qquad (1)$$

$$w_{t,d} = (1 + \log(\mathrm{tf}_{t,d}) \text{ x } idf \qquad (2)$$

Term Frequency is combined with IDF by simply multiplying the values as shown by(2).Where wt, d is the term weight, tft, d is number of occurrences of a specific term in a document and idf is the inverse document frequency.

K-Means Algorithm

The k-means clustering is known to be thorough in clustering huge data sets. The k-means algorithms aims to partition a set of objects,based on the attributes/features, into k cluster,where k is predefined and user-defined constant. The key point is to define k centroid, one for each cluster. The centroid of a cluster is formed in such a way that it is nearly related (in terms of similar function; similarity can be measured by using different methods such as cosine similarity, Euclidean distance, Extended Jac card) to all objects in that cluster.

Basic K-Means Algorithm

1. Choose k number of clusters to be determined
2. Choose k objects randomly as the initial cluster center.
3. Repeat .

Stemming

Many of the stretched out words in the English language generally fail to provide proper meaning in the given scenario and also they increases the computing 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 replaced with y).

Stop Word

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 minimize the overhead of processing.

Gaussian Distribution

a) This step actually adds the more words into the clusters based on the Gaussian distribution which is represented by Gaussian kernel equation.

Gaussian Kernel Equation

- (y-µ)2
2σ 2
P(y)= 1 e ------------------------------(2)
σ√2π

where,

µ= mean of distribution
σ2 = variance of distribution
y= continuous variable
P(y)= probability of y

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

Step 1: Generate N solutions xi(i = 1, 2, … ,n) from [αi, bi], using the uniform design technique.
Step 2: Assess the fitness of all individuals at first of the population and retain the appropriate results.
Step 3: Order the population by fitness in descending sorting and choose the desirable m individuals (m ≤ N).
Step 4: Resolve the generated m individuals information and calculate the mean *ui*^ and variance *σi*^ of each variable.
Step 5: Sample N new solutions according to formula (2).
Step 6: If the given stopping condition (up to the required variations n max) is not met, go to step 2.

Comparison with existing system

DISADVANTAGES OF EXISTING SYSTEM(ADA BOOST)	ADVANTAGES OF PROPOSED SYSTEM (FUZZY ANN &BAYESIAN PROBABILITY)
Use only labeled data.	Random(unlabeled) type of data is acceptable.
Does not form classification rule.	Forms classification rule.
Does not deal with class imbalance problems in minority class for misclassified data.	Helpful todeal with class imbalance problems in minority classes for misclassified data.

III. PROPOSED SYSTEM

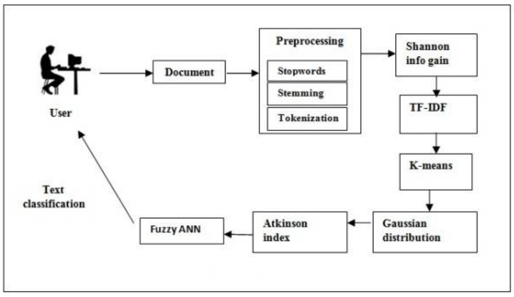

Fig : System Architecture

Here, we input an text document for classification output. The input document is pre-processed using stopwords, stemming and tokenization techniques. Further the semi-supervised algorithms like Shannon-info gain gain's required important data from the given output text document. The gained information is then calculated for its term frequency and inverse document frequency. The calculated term frequencies are clustered and distributed by Gaussian technique. Then the distributed data is indexed using atkinson index .The formed index is further implied as an input for Fuzzy ANN algorithm to get the required output of classified text.

PROCESS SUMMARY

1] Unlabelled Reuters data is taken as input data by user. The 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 stopword, stemming and tokenization techniques.
3] Shanon-info gain algorithm is used to gain very important data within the input document after the sorting pre-processing technique.
4] The important data within the document is calculated for its term frequency and inverse document frequency.
5] Calculated frequencies are clustered by using k-means clustering algorithm.
6] The clustered data is then distributed by Gaussian distribution technique.
7] Atkinson index technique is used to index the distributed data of the input data.
8] Lastly Fuzzy ANN algorithm is implemented on the index formed in the previous process for the desired output of classified text.

IV. CONCLUSION AND FUTURE WORK

The proposed system uses Ada-Boost technique which results to training error and normalization factor. Also the technique results to classification loss on target examples.[1] The system used is fuzzy L_R type number which only can classify Persian language text document. Also evaluates precision recall parameters.[2][3].Our system tries to overcome all these factors by using probability techniques. The experimental results to best performance, time efficient and accuracy. In future ,we will try to work on the domain of image classification.

REFERENCES

[1]Semi-Supervised Text ClassificationWith 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, Securityand 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.

[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.