Enhancing Binary Classification by Modeling Uncertaion Boundary in Three Way Decision using Multi Document Classification

N.Murugan,
P.G.Student ,Dept. of Computer Science and Engineering,
Anna university(BIT)Campus Trichy

ABSTRACT — *Text classification is a process of classifying documents into predefined categories through different classifiers learned from labelled or unlabeled training samples. The binary text classification attempt to find a more effective way to separate relevant texts from a large dataset. The current text classifiers cannot explain the decision boundary between positive and negative objects due to the uncertainties caused by text feature selection and the knowledge learning process. This paper proposes a three-way decision model for dealing with the uncertain boundary to improve the binary text classification performance based on the rough set techniques and centroid solution. Its ultimate aim is to make us understand the uncertain boundary through classifying the training samples into three regions as the positive, boundary and negative regions by two main boundary vectors. The four decision rules are proposed from the training process and applied to the incoming documents for more classification. A large number of text have been conducted based on the standard data sets RCV1.The proposed model has significantly improved the performance of binary text classification in term of measure and accuracy area compared with six other popular baseline models.*

keywords—*Uncertain decision boundary, text classification, three-way decision, rough set, decision rule*

I. INTRODUCTION

The explosive growth of electronic text documents, text classification,one of the crucial technologies of information organization and information filtering, is becoming increasingly important and attracting extensive attention in related research areas in recent years. Text classification plays a key role in both organising and seeking the relevant information that is usually labeled as positive from huge data sets. It is the process of classifying documents into predefined different categories based on their relevance to a topic, a category or a user.There are many practical applications of text classification, such as in news, e- mails, web pages, academic papers, medical records, and customer reviews . A number of text classification techniques have been developed, including support vector machines (SVM), Naive Bayes (NB), Rocchio that some objects relevant to a class have been omitted (loss of recall).overload means that some objects assigned to a class are actually not relevant to that class (loss of precision).The probability-based NB classifier has been used for a long time. The dealing with text information it is difficult to calculate the probability of the relevance of a given set approximation of the vague concepts. This is the theoretical basis for this study and the real reason why the three-region partitioning strategy for the training set is proposed in this study to solve the problem of the uncertainty
Decision boundary.The pair of lower and upper approximations divide the universe of objects into three disjoint regions the lower approximation as POS The difference between the upper and lower approximations as BND and the complement of the upper approximation as NEG,The mentioned in the introduction the notion of three-way decisions of terms even a set of term-weight pairs to a class because of complicated relations between the terms. Neural networks have been frequently adopted for computer vision and speech recognition.

For text classification a number of deep learning and neural network techniques were developed to improve the efficiency of learning neural networks. The time complexity of text representation learning for multi-label text classification can be reduced to the length of the text. The training process or word embedding requires a lot of computational resources.

II. RELATED WORK

Binary text classification is an important research issue in the area of information organization and information filtering that focuses on analysis and prediction of a document's correlation to a user, a topic or a class It is difficult for a classic binary text classier to set a clear decision boundary due to much uncertainty possibly caused by the deficiency of the classic classification algorithms or the knowledge learning process of specific classifiers. For example, our greatest concern is how to reach an effective

representation of the classes, the decision boundary or the related objects that need to be classified The Many popular binary text classification models have been well developed, such as NB, decision trees, Rocchio, k-NN, learner with its efficiency and SVM as a probabilistic classifier, is based on the application of Bayes theorem in decision rules. It is popular because it uses the dispersion feature to represent documents with much compactness, and also because it is highly efficient in dealing with binary classification problem such as anti-spam email filtering NB requires features to be independent from each other, which may bring more efficiency for the computation but limits its applicability. Decision tree based is an open source Java implementation of the algorithm which employs the entropy measure as its splitting function and uses the attributes with the highest normalised information gain to make the decision.

III IDENTIFY RESEARCH AND COLLECT IDEA

The researchers have believed that a decision class can be approximated by a pair of definable sets The lower approximation which consists of those objects that certainly belong to the decision class, and the upper approximation which consists of those objects that only possibly belong to the decision class. The some objects of interest cannot be discerned as same or similar due to the granularity of knowledge as they are assumed to be represented by the limited available information about a vague concept cannot be characterized by the relatively certain information about their elements, but can be replaced by a pair of precise concepts, the lower and the upper approximation of the vague concepts. This is the theoretical basis for this study and the real reason why the three-region partitioning strategy for the training set is proposed in this study to solve the problem of the uncertainty decision boundary. The pair of lower and upper approximations divide the universe of objects into three disjoint regions the lower approximation as POS the difference between the upper and lower approximations as BND and the complement of the upper approximation as NEG As The mentioned in the introduction the notion of three-way decisions.

IV THREE WAY DECISIONS

The introduction the notion of three-way decisions introduced by initially intended as a means to interpret classification/decision rules induced in probabilistic rough sets.The theory allows a risk based way to understand the lower and upper approximations, or the probabilistic positive, boundary, and negative regions. Currently three way decisions are formally described in an approximation space, where the probabilistic theory is used to measure lower and upper approximations for risk-based decision making. As mentioned in the previous section we can use the expected loss define three regions POS and NEG and BND In this paper, we call BND the uncertain boundary for which includes documents that cannot be clearly classified into POS or NEG because of the shortage of relevant knowledge about the class.To calculate the expected lost function efficiently some constraints have been introduced for the risk parameters For example risk parameters were assumed to be

$$Odds(X,d) > \theta 1 \rightarrow a1 \text{……..(1)}$$

$$Odds(X,d) > \theta 2 \rightarrow a3 \text{.....(2)}$$

V.MODELLING UNCERTAIN DECISION BOUNDARY

In the area of text analysis, it is time consuming very high time complexity to work out probabilities because of the complicated relations between terms keywords. NB is one of the popular techniques for text classification. By using Bayes Rule the goal of a Bayes classifier is to calculate each document, where variables means the years people have provided some effective methods to calculate the probability of the relevance for a given term in a set of documents. It is a very hard to estimate the correct probability because of the complicated relations between terms, such as polysemy and synonymy. Some approximation approaches have been proposed to estimate by combining those only depending on assumptions of event space the multiple multinomial model any assumption for the event space on free text information cannot correctly simulate the true situation.In this paper we propose an alternative approach to synthesizing these probabilities The view each term probability with its frequency in a document as a data dimension for describing the document and then develop decision rules to decide the relevance of the document to a given class. For this motivation we need to extend three-way decisions to vector spaces first.

$$\frac{p(X/d)=p(X).p(d|X)}{p(d)} \text{……(3)}$$

$$\frac{p(\neg X|d)=p(\neg d|\neg X)}{P(d)} \text{……..(4)}$$

Then,we have

Odds(X,d)

$$= \frac{p(x|d)=p(x).p(D|x)}{p(d)} = \frac{p(x)}{p(\neg x)}. + \frac{p(d|x)}{p(d|\neg x)} \text{……..(5)}$$

vector of weights for selected terms. The weight of each frequency in document .The generates three regions POS, BND and NEG and takes Similar to works out the derived boundary vectors, and the worst case takes calculates the mean distances, and steps and calculates the thresholds. These steps take The weight of each term is and is term frequency in document generates three regions (POS, BND and NEG) and takes Similar to the works out the derived boundary vectors and the worst case takes to calculates the mean distances and the thresholds The time complexity of Algorithm is the number of features. It only takes several vector operation sum and distance for each incoming document.

VI OVERALL ALGORITHMS CENTROID BASED SOLUTION

The proposed three-way decision approach is implemented training and Algorithm for testing application. To make the training process simple discussions The first step in is to select top terms by using a probabilistic model. This step first the number of documents containing the number of positive documents containing term, the time complexity of the first step where is the average size of documents.

$$P(w|D) = \frac{r(w)+0.5/(|d^{+}-r(w)+0.5)}{(n(w)-r(w)+0.5)/(|D|-n(w)-|D^{+}|+r(w)+0.5)} \quad \ldots. \quad (6)$$

It also assigns empty to sets POS, BND and NEG. The time complexity of the first step is the average size of documents to calculates the centroids for both positive and negative documents.

VII DATA COLLECTIONS

There are two ways to assign a class to a document the content based approach based on subjects in the document and request-based approach based on relevance to a particular audience or user group. The Human experts assigned labels to documents where each document includes about 50 words in average after pre processing based on subjects and decided relevance to RCV1 documents where each document includes about 130 words in average after pre-processing for a topic based on the title of the topic and description and narratives.

TABLE

The Baselines Models and Their Algoritham

No	Algoritham Type	Classifier
1	Linear kernel function	libSVM
2	Optimizing classification	SVMperf
3	rules	J48
4	Decision tree	Naive Bayes
5	Probability based	IBk
	Nearest neighbours (kNN)	

VIII EXPERIMENTS AND EVALUATIONS

The actual application of binary text classifiers, we find that usually there is a relative small number of objects in positive classes as compared to negative classes because any information that users do not want is non-relevant information. The score is a very important indicator for a successful binary text classifier The while accuracy is not very suitable. The select two kinds of RCV1 an unbalanced data collection in which a positive class is relative small and a balanced data collection in which the size of a positive class is close to the size of the corresponding negative class to evaluate the proposed approach. Both data collections are composed of XML typed documents and are widely used in the areas of text analysis and text classification

IX Parameter Tuning

Based on the binary classification rules the region that Belongs to can be compared by its odd and even value two pairs of decision parameters. The proposed classifier then needs to determine the decision parameters by using two pairs of experimental parameters To simplify the process of deciding experimental parameters we first decide the first pair according to the training data set The considering the number of positive documents is normally less than the number of negative documents and use the central line to classify the boundary region. The set the value of the second pairs as the same as the first pair shows some examples for determining the first where we selected the third based on its performance in the training sets.

X Scalability

The Algorithm its time complexity is decided by the number of features. For the proposed model the performance variation is also evaluated with the change of features on the two data sets. The results are demonstrated by tables and figures, from which it is not difficult to find out the relationship between the various variables that we are concerned about such as the values of Accuracy and the number of features. The same conclusion has been reached from the results on both of the two different data sets.

RESULT AND DISCUSSION

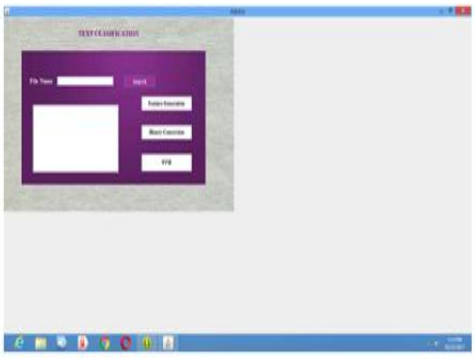

Fig No 1 sample classification for future generalization

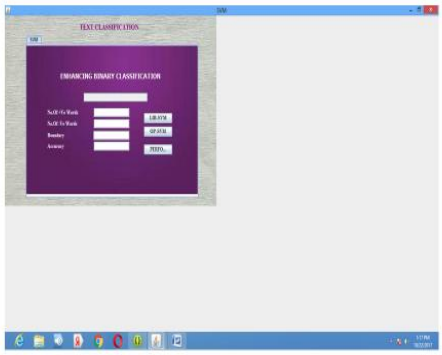

Fig No 2 sample classification for acuuracy and performance

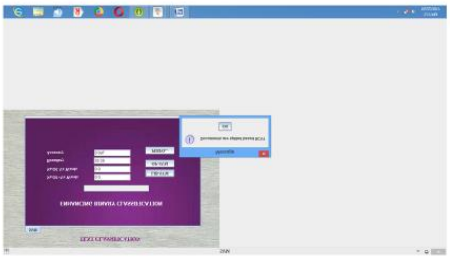

Fig No 3 enhancing for binary classification for splited baseddd based rcv1.

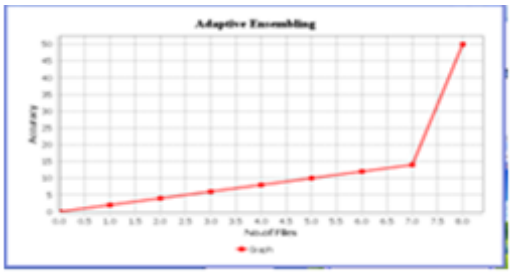

Fig No 4 Enhancing Performance with high accuracy

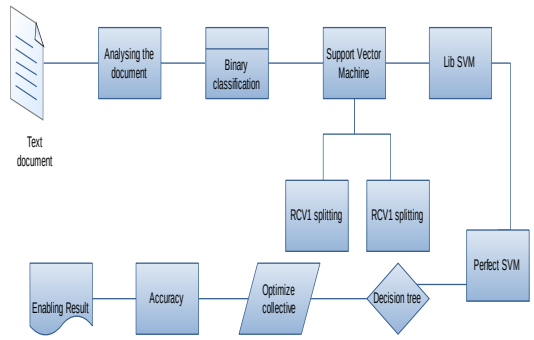

Fig No 5 Proposed System Architecture:

XII CONCLUSION

This paper for addressing problem of uncertain decision boundary to improve the performance of binary text classification The results show that the proposed model can significantly improve the performance of the binary text classification through this research the following conclusions can be made. This study has revealed that a satisfactory classifier can be implemented in an indirect way via an intermediate step of three region partitioning, and that the structure and properties of the boundary region obtained at the training stage can be applied to the incoming documents through the two pairs of boundary vectors both of which are based on the theoretical derivation and results. The study on the effect of the selected feature number on the other hand the time complexity for the proposed model to train the classifier is analyzed and the process is efficient. Both theoretical analysis of the proposed algorithms and the experimental results for the proposed classifier indicate that the proposed model can provide a promising direction for text classification. The contributions made by this study An efficient three-way decision model has been proposed to discover knowledge for representing uncertain information the boundary an effective classifier for binary text classification is proposed an in An effective upload the multi documents.

REFERENCES

[1] R. Y. Lau, P. D. Bruza, and D. Song, "Towards a belief-revision based adaptive and context-sensitive information retrieval system,"ACM Trans. Inf. Syst., vol. 26, no. 2, pp. 8.1–8.38, 2008.
[2] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proc. 16th ACMSIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 753– 762, 2007.
[3] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surveys, vol. 34, no. 1, pp. 1–47, 2009.
[4] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proc. 11th Conf. Uncertainty Artif. Intell.,, pp. 338–345,2008
[5] T. Joachims,"Transductive inference for text classification using support vector machines," in Proc. 16th Int. Conf. Mach. Learn,pp. 200–209,2010.
[6] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in Proc. 3rd IEEEInt. Conf. Data Mining,pp. 179–186,2013
[7] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Na€ıve Bayes," Expert Syst. Appl., vol. 36, no. 3,pp. 5432–5435, 2010.
[8] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining Text Data. Berlin, Germany: Springer, pp. 163–222, 2012.
[9] M. A. Bijaksana, Y. Li, and A. Algarni, "A pattern based two-stage text classifier," in Machine Learning and Data Mining in Pattern Recognition. Berlin, Germany: Springer, pp. 169–182, 2013.
[10] L. Zhang, Y. Li, C. Sun, and W. Nadee, "Rough set based approach to text classification," in Proc. IEEE/WIC/ACM Int. JointConf. Web Intell. Intell. Agent Technol, vol. 3, pp. 245–252,2013
[11] M. Haddoud, A. Mokhtari, T. Lecroq, and "Combining supervised term-weighting metrics for SVM textclassification with extended term representation," Knowl. Inf. Syst., vol. 49, no. 3, pp. 909–931, 2016.
[12] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surveys, vol. 34, no. 1, pp. 1–47, 2012.
[13] C. Manning, P. Raghavan, and H. Sch€utze, Introduction to Information Retrieval, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, pp.1-68,2008.