Volume-7 | Issue-4 | April-2018 | PRINT ISSN No 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

ADAPTIVE AND CONTEXT OF HOLISTIC HEALTH CARE SERVICES FOR RETRIVAL 0F INFORMATION

Information Technology

Shree Vaishnavi. R Department Of Information Technology- IV year Velammal Institute Of Technology

Deepika.R.C* Department Of Information Technology- IV year Velammal Institute Of Technology*Corresponding Author

Amritha.S Department Of Information Technology- IV year Velammal Institute Of Technology*Corresponding Author

ABSTRACT

Text classification is a process of classifying documents into predefined categories through different classifiers learned from labelled and unlabelled training samples. Many researchers who work on binary text classification attempt to a more effective way to separate relevant texts from a large data set. Here we proposes a three-way decision model for dealing with the uncertain boundary to improve the binary text classification performance based on the rough set technique and centroid solutions. Four discussion rules are proposed from the training process and applied to the incoming document for more precise classification. However, the current text classifiers cannot unambiguously describe the Decision boundary between the positive and negative objects because of uncertainties caused by text features selection and the knowledge learning process . The experimental results show that the usage of boundary vectors is very effective and efficient for dealing with uncertainties for the decision boundaries, and the proposed model has significantly improved the performance of binary text classification in term of F1 measure and AUC area compared with six other popular

KEYWORDS

I. INTRODUCTION

With 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. Text categorization(TC), the problem of assigning documents to predefined categories, is an active research area in information retrieval and machine learning. A wide range of supervised learning algorithms have been applied to this problem, using a training set of categorised documents to obtain an empirical mapping from arbitrary documents to relevant categories. This mapping is typically realized by assigning relevance scores to every document-category pair, and then thresholding on those scores to make binary decisions. Both the scoring method and the thresholding method used in categorizations systems can influence its results significantly. However, only the scoring algorithm(k-nearest neighbour, Naïve Bayes, multivariate regression, decision tree ,support vector machines ,neural networks, boosting etc) have been the major focus of research in the tc literature, while thresholding strategies were often briefly mentioned as a unimportant post-processing step. The implicit assumption was either that thresholding strategies do not make much difference in the performance of a classifier, or that finding the optimal thresholding strategy for any given classifier is trivial. Neither of the above assumptions is true. Optimal thresholding is trivial only of the classifier produces accurate probabilities p(cj/di) for all the categories(cj) and doucments, and I the optimization criterion is to minimize the global number of errors in the decisions made by the systems, and possibly some other conditions. Under those conditions, for a 2-category classification problem where a document belongs to one and only one category.

II.LITERATURE SURVEY

In this, past literature related to text classification will be reviewed and discussed. This review will cover the necessary theories, technologies and their application in the relevant areas, including text classification, text representation, text mining models, rough set decision, Bayesian decision, pseudo relevant feedback, feature selection and weighting and vector space and operations. The description and discussion of these theories, models and application provide comprehensive information about their current development and indicate their strengths and weakness in order to contribute to description and understanding of the research framework to be proposed in this study.

There are six parts to this literature review: Text classification and classic text classification algorithms. This part discusses the most popular text classifiers, and their characteristics, techniques and applications.

II.I.THREE WAY DECISION:

For a long period of time, researchers have believed that a decision class can be approximated by a pair of definable sets-the lower approximation and the upper approximation. According to pawlak's treatise, 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 them; therefore a vague concept cannot be characterized by the relatively certain information about their elements, but can be replaced by a pair of precise concepts; ie 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 portioning strategy for the training set is proposed in this study to solve the problem of the uncertainty decision boundary.

II.II. MODELLING UNCERTAIN DECISIONN BOUNDARY

In the area of text analysis, it is time consuming to work out probabilities P(X\d) because of the complicated relations between term. NB is the one of the popular technique for text classification. By using baye's rule, the goal of a bayes classifier is to calculate p(d\X)=p(w1,w2,w3m…wn/x) for each document, where variables wi=1 means features fi=d and wi=0 means fi+/d . over the years, people have provided some effective methods to calculate the probability of the relevance for a given term (p(wi\X)) in a set of documents. However, it is a very hard to estimate the correct probability p((w1,w2,w3,…wn)/X) because of the complicated relations between terms, such as polysemy and synonymy. Some approximation approaches have been proposed to estimate p(d/X) by combining those p(wi/X) only depending in assumption for the event space on the free text information cannot correctly simulate the true situation.

III.CLASSIFICATION III.I.PRINCIPLE:

Through the above theoretical derivation, it is known that for any incoming document u, we can assign a region to it by using odds(X,u) and the decision parameters p and n. Let F be the selected feature set, and $and\ \sim u\ (w1, w2, \ldots wk) be\ the\ selected$

III.II.THE ROCCHIO TECHNIQUE:

Rocchio is an early text classification method .In this method ,each feature value in the vector is the computed using the classic tf-idf scheme.let d b the whole set of training document in class cj building a Rocchio classifier is achieved by constructing a prototype vector cj for each class cj. $\mathrm{cj} = a_{\square} \square \frac{1}{|\mathrm{cj}|} + \sum_{d\epsilon cj}^{\infty} \left(\square_n \square \frac{\square}{L} + b_n \sin\frac{n\pi x}{L} \right) \alpha\ nd\ \beta\ are$

Parameter that adjust the relative impact of relevant and irrelevant training example recommends α=16 and β=4.

In classification, for each test document td, it uses the cosine similarity measure to compute the similarity of td with each prototype vector .the class whose prototype vector is more similar to td is assigned o td. The algorithm that uses Rocchio identify a set RN of reliable negative document from U .

IV.EXPERIMENTS

IV.I Performance measure

To evaluate the categories performance of Knm with various thresholiding strategies, I present the result in both micro-averaged and macro-averaged recall, precision an F1.Recall® is the proportion of correctly predicted YESes by the system among the true YESes for all the document category pairs given a dataset .precision(p) is the proportion of correctly predicated YESes among all the system predicated YESes, the F1measure is the harmonic average of recall and precision, define to be

$$F1=\frac{2pr}{p+r}$$

A more general notion for the F-measure is

Where parameter β is specified to adjust the relative weighting between recall and precision. When scores are micro-averaged, the binary decisions are collected in a joint pool and then the recall, precision and F1 values are computed from that pool. When the sources were macro-averaged, the recall , precision and F1 values for individual categories are computed first and then averaged over categories.

IV.II EXPERIMENTAL SETTINGS:

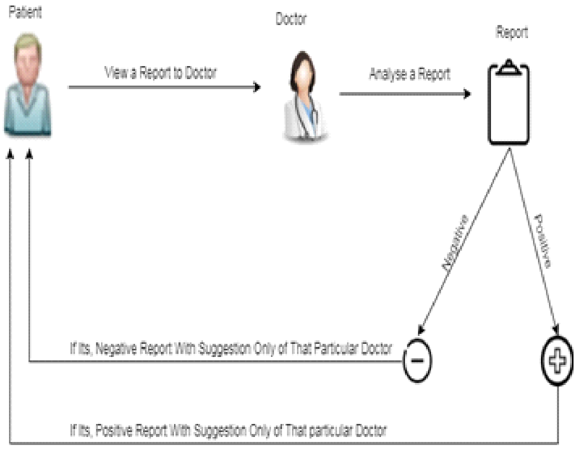
For the experiments in this paper, I used for standard KNN classifier for which detailed description and the parameter setting process were reported. Stop word removal, stemming and statistical feature selection were applied to documents in a pre-processing step, using the well. All the parameters for the thresholding strategies were tuned using hold-out validates sets; threshold increment unit was manually chosen for each corpus, to obtain sufficient number of recall-precision plots for an interpolated trade-off curve. For SCut, both SCutFBR.0 and SCutFBR.1 settings were examined on each corpus for the setting with a better performance in the validation condition, its performance on the corresponding test sets was reported for Scut.

IV.III EMPIRICIAL OBSERVATIONS:

The micro-averaged recall-precision curves for KNN with RCut, SCut and RTCut applied to three corpora. Those curves were obtained by thresholding at t=1,2,3.. in RCut, x=0.5,1,2,3.. in PCut and appropriate incremental thresholds inn RTCut to obtain high precision values at each threshold were computed and interpolated. SCut did not produce a curve, but a single point per data set instead. The “the break-even line” is drawn in those graphs as a reference line on which recall and precision have equal values and around which F1 scores are typically optimized. The performance differences in micro-averaged F1 on routers should be statistically significant, according to a related study on the same copora.

V.ARCHITECTURE DESIGN:

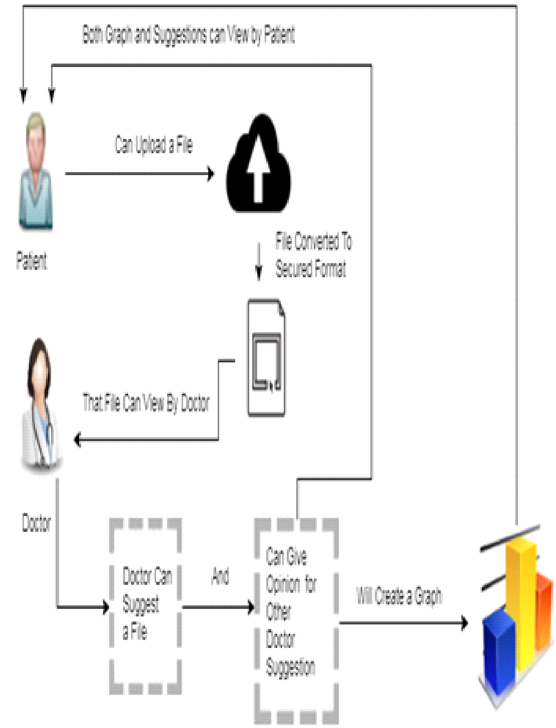
EXISTING SYSTEM:


Patient
Doctor
Report
View a Report to Doctor
Analyse a Report
Negative
Positive
If Its, Negative Report With Suggestion Only of That Particular Doctor
If Its, Positive Report With Suggestion Only of That particular Doctor

PROPOSED SYSTEM:

The necessary concepts, knowledge, theoretical derivation and proof, and overall algorithms that contribute to description and understanding of this research framework will be presented in the two subsequent sections.

The effectiveness and benefits of the proposed model will be demonstrated and evaluated by comparing it with the state-of-the-art baseline models. A large number of experiments have been conducted based on the proposed approach for text classification using the two standard data sets: RCV1 and R21578, including the comparative analysis between the proposed model and six baseline models, and in-depth analysis of the progressive improvement by the proposed model and the effectiveness of the derived boundary vectors.


Both Graph and Suggestions can View by Patient
Can Upload a File
Patient
File Converted To Secured Format
That File Can View By Doctor
Doctor
Doctor Can Suggest a File
And
Can Give Opinion for Other Doctor Suggestion
Will Create a Graph

V1.EXPERIMENTS AND EVALUATION:

When we examine 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.

VI.I. DATA COLLECTIONS:

Traditionally, there are two ways to assign a class to a document; the content –based approach(based on subjects in the document), request-based approach(based on relevance to a particular audience or user group).

VI.II. BASELINE METHODS:

When the baseline models were running, their parameters Were manually adjusted to help obtain the maximum Performance in testing sets. The model selection criteria is that the model either uses a similar key technology with the proposed model, such as Rocchio, or is regarded as the most popular models with the first-class performance, such as the other five baseline models.

VII.RESULTS:

In our experiments, the two pairs of boundary vectors of CP ; C~N (main) and B~P ; B~N (derived), and the decision rules are worked out through training process. Then, the documents in the incoming document sets, which are the testing corpus instead, are to be vector zed, with the decision Rules applied to reach the purpose of binary classification. In order to obtain the satisfactory effect, the Weighting coefficients of the radius have been adjusted for different data sets in the calculation of _P and _N, where P = 1:095; N = 3:177 for RCV1; P = 3:095; N = 4:177 for R56CO. The classification performance on both

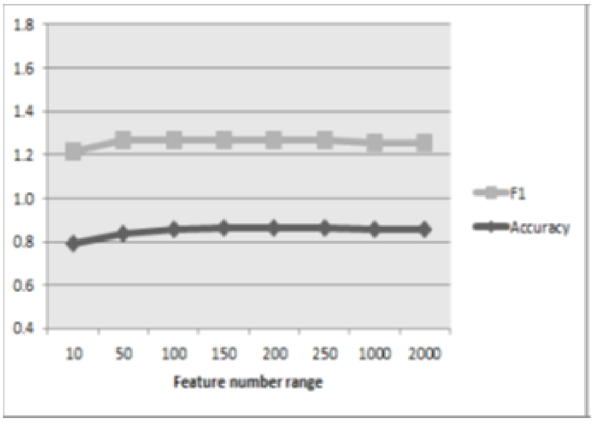
F1 measures and AUC is significantly improved for both RCV1 and R56CO data sets. The comparison between the proposed model TWDUB (Three-Way Decision for Uncertain Boundary) and other six baseline models has been completed using Three measures of F1, Accuracy and AUC. The results of the proposed model and six baseline models.

The RCV1, it is found that the proposed model has achieved the best performance. The percentage of Improvement over F1 and AUC is 9:5% and 5:0%, respectively, comparing with the best one of baseline models. The Accuracy value obtained by the proposed model exceeds that of the libSVM model, which has the highest Accuracy value among all the baseline models. Compared with the libSVM, the F1 score and AUC value of TWDUB have been significantly improved by 22:6% and 7:7%, respectively. For R56CO, it is found that the proposed model has achieved the best performance on both F1 score and AUC value. The percentage of improvement over F1 and AUC is 10:9% and 7:5%, respectively, comparing with the best one of baseline models. The SVMperf model gets the best Accuracy value, and the proposed model and libSVM on Accuracy are much closed to SVMperf. SVMperf also gets better performance on F1 than libSVM. Compared with the SVMperf, the F1 score and AUC value of TWDUB have been significantly improved by 11:1% and 7:7%, respectively.

VIII. ANALYSIS:

SCALABILITY:

The 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 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 or F1, and the number of features. The same conclusion has been reached from the results on both of the two different data sets including RCV1 and R56CO. The result of RCV1 and for the result of R56CO clearly show that the curved lines are relatively flat when the feature number increases to around 150 (the middle part of the curved line), with an obviously lower value at the start point of feature number, 10-100, and a constant or a bit lower value after 150 (for RCV1), or a bit higher or remaining unchanged value at a high feature number, 1000-2000 (for R56CO). If the features are reduced too much, the objects which are represented by the features will be seriously distorted or deviated, so as to cause a big decrease of the tested results by the same proposed model with the same parameter settings. It is also not meaningful if the feature number is over a certain scope as the time complexity will be increased exponentially, which is just the reason why the feature number was set to 150 in the proposed approach of this research.



IX. CONCLUSION

This paper proposed an innovative three-way decision approach for addressing the problem of uncertain decision boundary to improve the performance of binary text classification. The experimental results show that the propose model can significantly improve the performance of the binary text classification in terms of F1 and AUC, and can achieve a high Accuracy compared with other six baseline models. 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 learned boundary vectors, both of which are based on the theoretical derivation and experimental results. The study on the effect of the selected feature number shows that the performance is decreased significantly when IEEE Transactions on Knowledge and Data Engineering, volume: 29,No:7,July 1 2017 The feature number is lower than 100, remains relatively Stable when the feature number is around 150 (a suitable Size to control the efficiency). On the other hand, the time Complexity for the proposed model to train the classifier is Analysed 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 Are summarized as follows: _ An efficient three-way decision model has been proposed to discover knowledge (the boundary vectors) for representing uncertain information (the boundary) between two classes. An effective classifier for binary text classification is Proposed in an indirect way by adding a transitional step. In other words, the binary text classification is completed through two transformations: 'two-way to three-way', and then 'three-way to two-way'. The proposed method uses only a centroid for each class. In the future, we will extend the proposed method for multilabel document classification or multi-class classification.

REFERENCES

[1] R. Y. Lau, P. D. Bruza, and D. Song, “Towards a belief revision based adaptive and context-sensitive information retrieval system,” ACM Transactions on Information Systems, 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 Proceedings of KDD'10, 2010, pp. 753–762.
[3] F. Sebastiani, “Machine learning in automated text categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.
[4] G. H. John and P. Langley, “Estimating continuous distributions in bayesian classifiers,” in Proceedings of 11th conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.
[5] T. Joachims, “Transductive inference for text classification using support vector machines,” in ICML, 1999, pp. 200–209.
[6] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, “Building text classifiers using positive and unlabelled examples,” in Proceedings of ICDM'03, 2003, pp. 179–186.
[7] J. Chen, H. Huang, S. Tian, and Y. Qu, “Feature selection for text classification with na¨ıve bayes,” Expert Systems with Applications, vol. 36, no. 3, pp. 5432–5435, 2009.
[8] C. C. Aggarwal and C. Zhai, “A survey of text classification algorithms,” in Mining text data Springer, 2012, pp. 163–222.
[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, Springer, 2013, pp. 169–182.
[10] L. Zhang, Y. Li, C. Sun, and W. Nadee, “Rough set based approach to text classification,” in 2013 IEEE/WIC/ACM International Joint Conferences, vol. 3, 2013, pp. 245–252.