MEDLINE DOCUMENT CLASSIFICATION MODEL

S.sagar Imambi*

T.Sudha**

Abstract

Medline is online repository of medical literature As the natural language documents are unstructured, finding relevant features are very complex problem, and non linear distribution of documents increases the problem to find the decision surface that separates the data. We proposed a prototype based classifier to resolve these problems We empirically tested the model with Medline database and bench mark dataset Reuters 21875. As Medline data contain more complex unstructured data , it need better tools to extract information from the Medline and to categorize them. With our proposed algorithm we achieved nearly 90% of accuracy to classify the documents.

Keywords— Classification, Online repositories -Medline documents ,Text mining

* Asst Professor,TJPS College, Tirupathi

** Professor,SPMVV,Tirupathi

1.MEDLINE :

MEDLINE is the largest component of PubMed, the freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine. Approximately 5,400 journals published in the United States and more than 80 other countries have been selected and are currently indexed for MEDLINE. (http://www.ncbi.nlm.nih.gov/pubmed). PubMed Comprises more that 22 million biomedical literature from Medline, life science journals and online books. Categorization of Medline documents becomes a challenge as the repository includes unstructured text data. Identifying the concept terms from the documents to classify them is a very complex problem. As the natural language documents are unstructured, finding relevant features are very complex problem, and non linear distribution of documents increases the problem to find the decision surface that separates the data. We proposed a prototype based classifier to resolve these problems We empirically tested the model with Medline database and bench mark dataset Reuters 21875.

2. Survey of literature

PubMed and Medline documents provide very flexible resource for researchers. Padmini Srinivasan (2002) presented a system which is developed for the extraction of pairs of concepts from Medline dataset and to find the association between them. Padmini Srinivasan et al (2004) also developed Open discovery algorithm to uncovering implicit information from the Medline documents. They used this method to investigate the potential of turmeric or Curcumin Longa and identifies a ranked list of problems. Go-tag (M.Ghanem 2005 , NEWS-ML(Kam-Fai Wong 2002), MedMESH (P.Kankar 2002), IndexFinder (Qingghua Zou ,2003) are the novel architectures developed to analyze Medline document collections. Thuy.T.T. Nguyen et al (2007) presented an improvement on k-mean algorithm k-mix and allows its applications to the mixer of attribute types found in the cardio vascular domain.

Document classification has been studied intensively because of its wide applicability in areas such as web mining, information retrieval.(Aurangzeb Khan et al 2010). Automated classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, which includes grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted

content for Internet browsers. (Sebastian, 2002). Recently, numerous research activities have been conducted in the field of document classification, particularly applying in spam filtering (S.J.Delany 2005, P.Cunningham,2003), emails categorization (B.Kamens 2005, S.Sagar Imambi et al 2008), web site classification (M.I.Devi 2008), formation of knowledge repositories , ontology mapping (X.Su 2002), digital library (S.Sagar Imambi et al 2010) More recently, classification is also applied for browsing a collection of documents or organizing the query results.

3. Prototype based Classifier:

Main significant challenges of text classification are difficulty to capture abstract concepts(features) of natural language documents, quantify /weight these features, selecting relevant features for each class, choosing proper representation model of documents, availability of sufficient training documents, selection of classifier for large sets of closely related classes , Building classifier for nonlinearly distributed documents

. The traditional classification techniques like SVM, Decision tree and Baye's have not been producing better results as they are not able to generate the accurate decision plane to classify the documents. Having a better classification accuracy on biomedical literature can make an impact on the diagnosis of very complex and rare cases.

We have a set of predefined categories and a set of documents .For each category, the document set is partitioned into two mutually exclusive sets of relevant and irrelevant documents. The goal of a text classification system is to determine whether a given document belongs to any of the predefined categories. The document can belong to any one of the predefined categories.

There are 4 steps in construction of classifier.

1. Preprocess the collected data, remove the stop words and generate stemmed mesh vocabulary.
2. Terms are given weightage, so that we can find the relevance of each term. Basing on the relevance of term, feature is selected.

3. All the abstracts are represented in vector space model with reduceddimensionality. And Prototype vectors are generated for each class .The classification model was build by using the prototype vector.

4. 66% of data is used as training data to learn the model and remaining is used as test data To evaluate a text classification system. Harmonic mean of precision and recall (F-Measure) is one of the performance measures.

This measure combines recall and precision in the following way:

Recall =number of correct positive predictions/number of positive examples

Precision =number of correct positive predictions/number of positive predictions

F1 =2 *Recall *Precision /(Recall + Precision). If F-measure is near to 1 performance is good , if it is nearer to zero then classifier is not suitable.

4.Datsset

We collected the documents related to diabetic complications. Diabetes is a disease in which blood glucose levels are above normal. The number of people with diabetes is increasing in proportion to population growth. Ageing, stress due to urbanization and increasing prevalence of obesity and physical inactivity are some of the major contributors to diabetes. Assuming that age-specific prevalence remains constant, the number of people with diabetes in the world is expected to approximately double between 2000 and 2030, based solely upon demographic changes. People must be aware of diabetic symptoms and causes , so that prevalence of diabetes can be organized. Long term Diabetes can damage the kidney and leads to kidney failure. Damage to the retina from diabetes is a leading cause of blindness. Diabetes damage the nerves in the automatic nervous system and lead to paralysis Diabetes predisposes people to high blood pressure and high cholesterol and triglyceride levels. These conditions independently and together with hyperglycemia increase the risk of heart disease, kidney disease, and other blood .

We collected 8000 and 10,000 documents from Medline related to diabetes complications. We also collected 2000 and 6000 documents from Reuters data sets.

5. ExperimentalResult:

We applied the algorithm on the above mentioned data sets. And tested the performance. We also compared our algorithm with Navie Bayes classification model. The table1 shows the result of Bayes classification.Table 2 shows the results of algorithm.

Validation Measure	Medline data sets		Public data set	
	Med7	Med8	Reuter1	Reuter2
Accuracy	73.10	86.43	63.45	53.74
Precision	0.782	0.88	0.773	0.541
Recall	0.731	0.864	0.635	0.537
F-measure	0.742	0.869	0.687	0.532

Table 1. Performance measure of Navie Bayes Classification method

Data set	Precision	Recall	F-measure	Accuracy
Med7	0.92205	0.82902	0.836325	99.948
Med8	0.98388721	0.990239	0.9867911	98.3328
Reuter1	0.978088	0.996577	0.987081	97.56
Reuter2	0.956074	0.961508	0.958258	95.23

Table 2 Performance measures of GRPB Classifier.

We can see significant improvement of accuracy and F-measure. Med7 is classified with highest accuracy and precision. Reuter 2 shows lowest accuracy i.e., 95%. Accuracy is improved 20% with our algorithm.

Conclusion

As Medline data contain more complex unstructured data , it need better tools to extract information from the Medline and to categorize them. With our proposed algorithm we achieved nearly 90% of accuracy to classify the documents.

References:

B. Kamens(2005),” Bayesian filtering: Beyond binary classification” . Fog Creek Software, Inc., 2005.

D.V. Chandra Shekar and S.Sagar Imambi(2008), Classifying and Identifying of Threats in E-mails – Using data mining techniques- Proceedings of International Multi conference of Engineers and Computer Scienes-2008, Hongkong 19-21-March,2008 ,PP562-566

Kam-Fai Wong et al(2003), " Improving Document Clustering by Utilizing Meta-Data", Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, pp. 109-115.

M. Ghanem et al (2005) ,” GoTag: A Case Study in Using a Shared UK e-Science Infrastructure for the Automatic Annotation of Medline documents” , I4th UK e-Science All-Hands Conference AHM .

M.I. Devi, R. Rajaram, and K. Selvakuberan(2008) , “Generating best features for web page classification”, Webology, Vol. 5, No. 1,Article 52.

P. Kankar et al (2002) ," MedMeSH Summarizer: Text Mining for Gene Clusters", In the Proceedings of the Second SIAM International Conference on Data Mining.

Padmini Srinivasan et al (2002), " Hierarchical text categorization using neural networks” , Information Retrieval, Vol.5 pp.87–118.

Padmini Srinivasan et al(2003), "Categorization of Sentence Types in Medical Abstracts", AMIA 2003 Symposium Proceedings .

Padmini Srinivasan et al(2004), “ Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases” , Workshop: Biolink ,Linking Biological Literature, Ontologies and Databases, pp. 33-40.

. Qinghua Zou et al(2003) "IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing",Proceddings of Symposium AMIA -03,pp-763-767.

. S.J. Delany, P. Cunningham, and L. Coyle(2005), “An Assessment Of Case-Based Reasoning For Spam Filtering”, Artificial Intelligence Review Journal, Vol. 24, No. 3-4, pp. 359-378.

. S.Sagar Imambi, T.Sudha (2010)- A Unified frame work for searching Digital libraries Using Document Clustering –International Journal of Computational Mathematical ideas Vol 2-No1-(2010) ,pp 28-32

. Sebastiani, F.(2002), " Machine learning in automated text categorization", ACM Computing Surveys 1(34) (2002) pp.1-47.

. Thuy.T.T. Nguyen and Darryl. N. Davis(2007) ," A Clustering Algorithm for Predicting Cardio Vascular Risk", Proceedings of the World Congress on Engineering , Vol 1.

. X. Su(2002)," A text categorization perspective for ontology mapping,", Department of Computer and Information Science, Norweigian University of Science and Technology.