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. 3, Issue 12, December 2015

An Enhanced Text Mining Classification Model using EM Algorithm with Kernel for Drugs based on Data Reviews

Jyotsna Pulipati[1], Dr. P. Govindarajulu[2]

Research Scholar, Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India[1]

Professor, Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India[2]

ABSTRACT: Text mining is a hot research area now a days. With fast rising of its advanced engineering, business papers, e-mail and all data stored in electronic form so the large amount of data in and extracting a task relevant data from the large document is complex task. Here we look some significant text classification techniques which is basically use to classify the text document into predefined class. In theory, it is possible for a chemical to be active for multiple therapeutic indications as drugs. Yet in practice, several obstacles can impair the development of a potential new usage. The first factor to handle is serendipity. But based on the text classification techniques we have proposed a method called feature selection along with classification and clustering based on Gini Index. We have used text mining with EM algorithm with kernel based function to perform efficient classification on drugs data. The experimental results on drugs data shows that EM Algorithm with Kernel is capable to discover improved aspects than other frequent approaches, when considered with mean point wise reciprocal information and classification accuracy.

KEYWORDS: Data mining, Text Mining, Classification, Clustering with Kernel, Drugs data.

I. INTRODUCTION

A. Text Mining

Text mining is an information based procedure that uses logical tools to obtain meaningful information from the natural language text. The information is subsequent from the text by identifying and detecting significant patterns from unknown textual data. More specifically, text mining deals with extracting useful patterns from unstructured data rather than structured data. Text mining is useful to maintain the processes of:

- Monitoring protection by study of text from online sources, such as internet news and blogs.
- Enterprise business intelligence for discovering the competitors.
- Mining information for biomedical applications.

However, the text mining techniques come across the issue of correct illustration of concepts that may lead to mining of inaccurate structures from the text. Moreover, text mining approaches are also inadequate in properly classifying synonyms from huge documents containing the formless text[2,5]. The text mining techniques used in patent analysis are largely based on natural language processing, property function approaches, rule, neural networks and semantic based approaches.

- Nearest Neighbor Classifier.
- Bayesian Classifier.
- Support Vector Machine.
- Expectation Maximization Classifier.
- Association based Classification.
- Centroid based Classification
- Classification Using Neural Network

B. Text Classification

Text classification is the procedure of classifying documents into predefined categories based on their content. And also it is the task of assigning predefined categories to free text documents. Text classification is used in several fields

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. 3, Issue 12, December 2015

such as patient information in healthcare organizations are often indexed from several aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on.

C. Clustering

Cluster analysis is the task of grouping a set of objects in such a way that objects in the similar group are more related to each other than to those in new groups . It is a major task of retrieving information from data mining and a frequent techniques for numerical data analysis are used in many fields such as machine learning, image processing, pattern recognition and information retrieval. Work with clustering helps to modify data preprocessing and model parameters until the result achieves preferred properties.

II. LITERATURE REVIEW

Text mining comprises the intelligent automated analysis of textual data and aims for extraction of interesting facts and relationships and discovery of knowledge from large amounts of text. For this purpose, text mining employs techniques and algorithms from restraints such as data mining, information retrieval, statistics, mathematics, machine learning and natural language processing. Today, text mining is even used for emerging trend detection, policy-making processes, intelligence services, press monitoring to automatically detect breaking news, marketing, data protection, law enforcement and personalized advertising. As most information is currently stored as text, text mining is believed to have a high commercial potential value. Although the successful applications of text mining have been achieved in many areas, the challenges still remain. For instance, text mining usually lacks the deep and fully understanding of the literature and the information one needs is often not recorded in textual form. In many real-world problems, things can be illustrated by various sets of features. For instance, in systematic literature mining, both the textual content and the citation link between articles are often used in the information innovation process. In complex network analysis, we are given a set of multiple networks that share the same nodes but possess network specific links representing different types of relationships between nodes. Text mining helps biologist to automatically collect structured biomedical knowledge from large volumes of biological literature. Throughout the past few years, there was a flow of attention in routine investigation of the biomedical literature, ranging from the reserved approach of annotating and extracting keywords from text content[3,7].

Natural language processing

The Natural Language Processing is one of the text approach in text mining, that uses computational mechanisms to evaluate and characterize the textual information in documents. In patent analysis, the NLP has also been used for the conversion of the scientific information into uncomplicated language structures by extracting the grammatical structures from the textual data and creating the structural interaction among the components. The NLP based text mining approaches are generally categorized into:

1. Keyword based approaches.
2. Subject-Action-Object (SAO) based approaches.

Although keyword based text mining approaches are easy to execute, they require in illustration of significant scientific concepts and relations. The keyword based approach involves predefining keywords and key phrases that necessitate professional awareness. The SAO based text mining techniques are proficient for analyzing unstructured information by representing the relationships among key scientific mechanism. The patent documents are altered into SAO structures and every structure consists of a Subject (S), Action (A) and Object (O). The SAO structures are extracted honestly from the patent documents. But, the Natural Language Processing based approaches undergo from the issues of lexical and grammatical ambiguities and also require in representing the semantic relationships among the grammatical structures. The NLP based approaches have proved tremendously successful in processing huge documents containing enormous volumes of textual data[4,11].

III. SIGNIFICANCE OF DRUG REPOSITION

Drug repositioning is the identification of new therapeutic indications for known drugs. These drugs can either be approved and marketed compounds used daily in a clinical setting, or they can be drugs that have been shelved, namely molecules that did not succeed in clinical trials or for which projects have been discontinued for various reasons. In one sentence, drug repositioning can be defined as renewing failed drugs and expanding successful ones. One motivation behind drug repositioning is the possibility to further market and extend the application line or patent life of a drug, therefore increasing the revenue stream generated from it. Another aim is the treatment of rare or neglected diseases;

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. 3, Issue 12, December 2015

usually such conditions are difficult to address for financial reasons, yet there might exist some safe and active molecules already developed for other indications, deemed suitable for this scenario. A drug rarely keeps its original indication it gets reoriented throughout the years when more data becomes available and in vivo pharmacology better understood. Famous repurposing and discovery stories were mostly due to chance and unexpected results therefore it can be difficult to forecast any relevant opportunities [13].

Drug bank

The Drug Bank database is a unique bioinformatics and cheminformatics resource that combines exhaustive drug data with inclusive drug intention information. The approved drugs acting on proteins are extracted from the database and imported in the FTC knowledge base. In order to be selected, a compound must firstly be approved and secondly have a pharmacological action on at least one human protein target present in Uniport. The Drug bank actions are somehow structured and consistent: concepts such as inhibitor or agonist are reused throughout the database for example, yet they are not strictly formalized as a controlled vocabulary.

IV. PROPOSED FEATURE SELECTION FOR TEXT CLASSIFICATION

Prior to the classification assignment, one of the most primary responsibilities that desires to be proficient is that of document illustration and feature selection. Although feature selection is also attractive in other classification tasks, it is particularly essential in text classification due to the elevated dimensionality of text features and the subsistence of unrelated or noisy features. In general, text can be represented in two disconnect ways. The primary way is as a bag of words in which a document is represented as a set of words and mutually with their connected occurrence in the document. Such a representation is basically independent of the series of words in the group. The second method is to represent text instantly as strings, in which every document is a series of words. The majority text classification methods use the bag-of-words illustration because of its ease for classification purposes. In this research work we have discussed several methods which are used for feature selection in text classification.

Gini Index

One of the most frequent method for quantifying the discrimination level of a feature is useful for an evaluation is known as the Gini-Index. Let $p_1(w)...p_k(w)$ be the part of class label existence of the k dissimilar classes for the word w. In further words, $p_i(w)$ is the restricted probability that a document belongs to class i, given the truth that it contains the word w. consequently:

$$\sum_{i=1}^{k} p_i(w) = 1$$

Then the Gini-Index for the word w, represented by G(w) is defined as follows:

$$G(w) = \sum_{i=1}^{k} 2 * p_i(w)$$

The importance of the Gini-Index G(w) constantly lies in the range (1/k, 1). Elevated principles of the Gini-Index G(w) symbolize indicate a better discriminative control of the word w. For instance, when all documents contain word w belong to a particular class, the value of G(w) is 1. Alternatively, when documents containing word w are uniformly scattered among the k different classes, the value of G(w) is 1/k.

Let $P_1...P_k$ stands for the overall distributions of the documents in different classes. Then, we establish the normalized probability value $p_i(w)$ as follows:

$$p_i(w) = \frac{p_i(w)/p_i}{\sum_{j=1}^{k} p_j(w)/p_j}$$

V. IMPLEMENTATION WITH STRING KERNEL

A. Kernel based EM Algorithm

A proficient EM Algorithm is developed for parameter estimation. This section covers string kernels which are methods dealing with text directly, and not anymore with intermediary depiction like word document matrices. Kernel-

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. 3, Issue 12, December 2015

based clustering methods like kernel k-means, use an embedded mapping of the input data into an elevated dimensional feature space defined by a kernel function k.

$$k(x,y) = (\Phi(x), \Phi(y))$$

with the projection Φ: X → H from the input domain X to the feature space H. In other words this is a function returning the inner product (Φ(x), Φ(y)) between the images of two data points x and y in the feature space. All computational tasks can be performed in the feature space if one can find a formulation so that the data points only emerge inside inner products. This is frequently referred to as the kernel trick and is computationally simpler than explicitly projecting x and y into the feature space H. The major improvement is that the kernel calculation is by extreme computation expensive than operating frankly in the feature space. This allow one to exertion with high-dimensional spaces, including normal texts, normally consisting of several thousand word proportions. String kernels are defined as comparison measure between two sequences of characters x and y. The standard form of string kernels is given by the equation

$$k(x,y) = \sum_s \lambda_s \delta_{s,t} = \sum_s num_s\,(x) num_s\,(y) \lambda_s$$

where Σ represents the set of all strings and $num_s(x)$ denotes the number of occurrences of s in x and λ_s is a weight or molder factor which can be preferred to be set for all substrings or can be set to a dissimilar values for every substring. This basic sign includes a great number of exceptional cases are $\lambda_s = 0$ for all s>n, that is comparing all substrings of length less than n, often called full string kernel. the case $\lambda_s = 0$ for all $s \neq n$ is frequently referred to as string kernel. The length of s, u is a subsequence of s, if there exist indices $i=(i_1,..,i_u)$, with $1 < i_1 < \cdots < i_u < s$. Which can be used for dynamic program aspects to speed up calculation considerably. Assume that, we have an opinion problem in which we have a training set {x (1), . . . , x(m)} consisting of m autonomous examples. We wish to fit the parameters of a model p(x, z) to the data, where the probability is given by

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x;\theta) = \sum_{i=1}^{m} log \sum_z p(x,z;\theta)$$

But, unambiguously finding the maximum probability estimates of the parameters θ may be hard. Here, the z (i) 's are the dormant accidental variables and it is frequently the case that if the z (i) 's were experiential, then maximum probability opinion would be simple. In such a situation, the EM algorithm gives an proficient method for highest probability opinion. Maximizing ℓ(θ) explicitly might be difficult and our approach will be to instead repetitively build a lower-bound on ℓ (E-step), and then optimize that lower-bound (M-step). Repetitively carrying out these two steps gives us the EM algorithm, which is as follows:

E-Step for each i, set $$Qi(z(i)) = p(z^{(i)}|x^{(i)};\Theta)$$

M-Step set $$\Theta = argmax \sum_i \sum_{z(i)} Q_i(z^{(i)} log \frac{p(x^{(i)},z^{(i)};\Theta)}{Q_{(i)}(z^{(i)})}$$

suppose $\theta^{(t)}$ and $\theta^{(t+1)}$ are the parameters from two consecutive iterations of EM with string kernel. We prove that $\ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)})$, which shows EM always monotonically improves the log-likelihood. The key to showing this result lies in our choice of the Q_i's. Exclusively, on the iteration of EM with string kernel in which the parameters had started out as $\theta^{(t)}$, we would have chosen $Q_i^{(t)}(z^{(i)}) = p(z^{(i)}|x^{(i)};\theta^{(t)})$.

B. Flow Diagrams

A graphical tool used to illustrate and examine the instant of data through a system guide or automated including the procedure, stores of data and delays in the system. Data Flow Diagrams are the essential tool and the origin from which other components are developed. The transformation of data from input to output, through processes, may be described reasonably and separately of the objective components associated with the system. The Data Flow Diagram is also know as a data flow graph or a bubble chart. Data Flow Diagrams are the model of the proposed classification. We clearly shown the requirements on which the new system built. Afterward, throughout design activity this is taken as the origin for drawing the system's structure charts. The fundamental information used to create a Data Flow Diagram's are as follows:

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. 3, Issue 12, December 2015

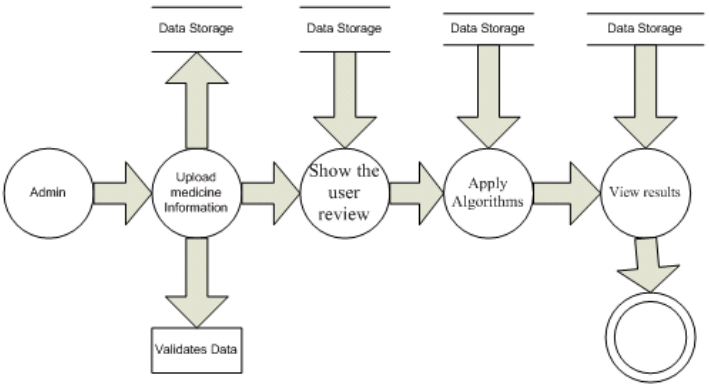

Fig.1: Admin DFD

C. Implementation System

The more information on a drug's molecular targets and their physiological roles, the more opportunities exist to re-orient a drug into doing something new. More importantly, these descriptors can also be used to identify outliers to the rule, in other words, drugs that are functionally similar yet clinically used for different indications. This set of drug pairs was defined as the drug repositioning opportunities. We have implemented our proposed kernel based EM algorithm with java code using mysql data bases to build a platform for analysing various drugs information based on the reviews. But here we developed an environment for text classification analysis of the proposed research work. The following login pages indicates the information about the administration activities which are performed at the beginning.

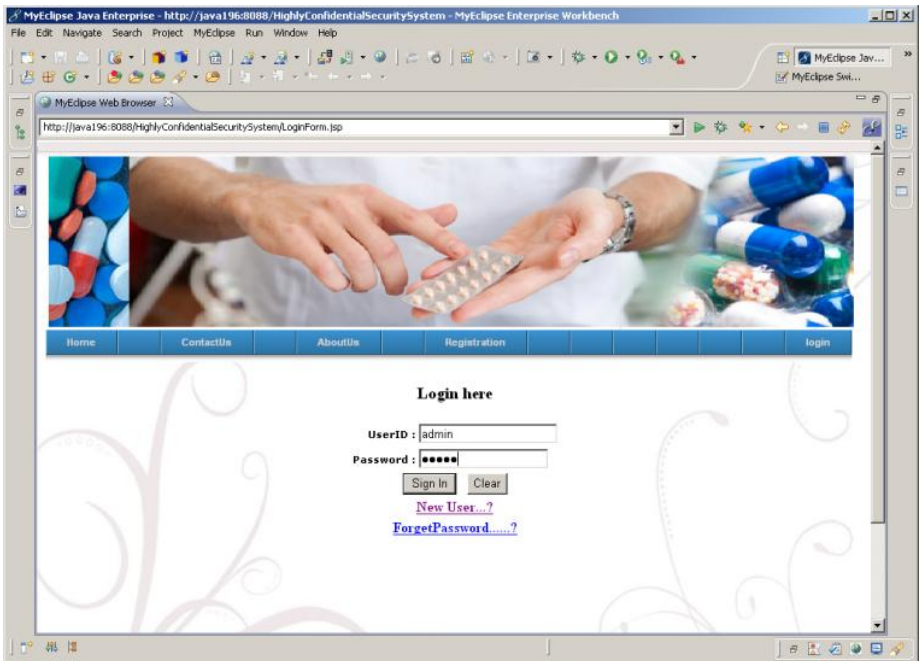

Fig.2: Admin Login page

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. 3, Issue 12, December 2015

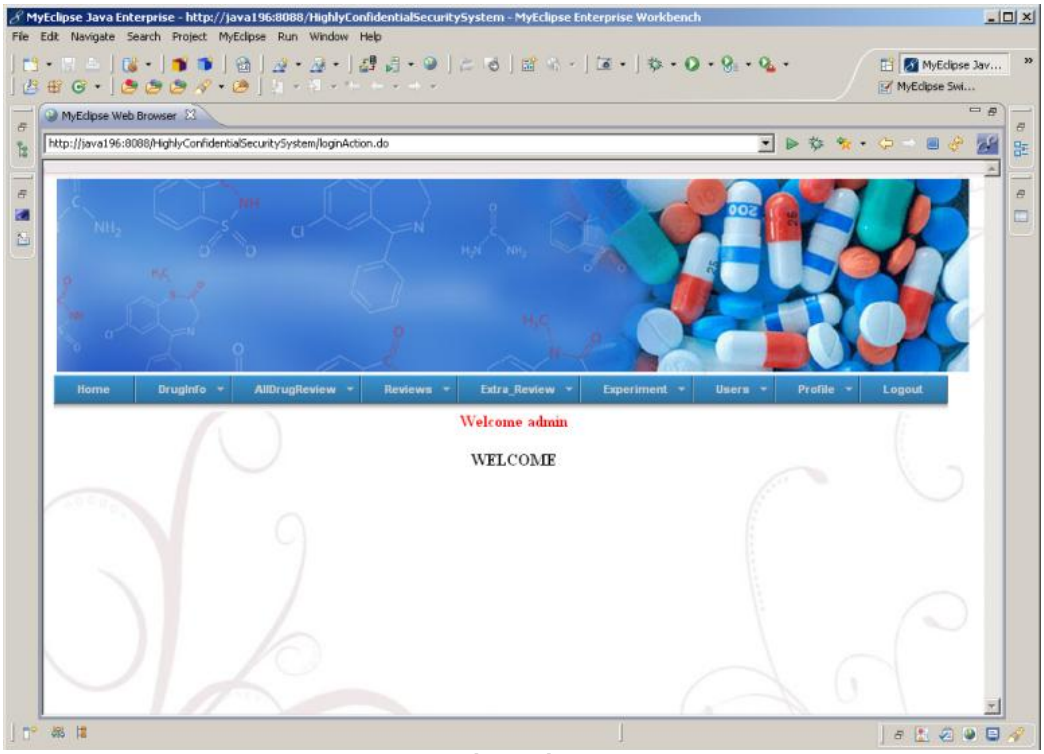

Fig.3: Admin Welcome page

VI. CONCLUSION AND FUTURE WORK

Text Data mining is also known as Knowledge Discovery in Text (KDT). Generally the process of extracting interesting knowledge and information from unstructured text. It is a young interdisciplinary area which draws on information recovery, machine learning, statistics and computational linguistics. The problem of Knowledge Discovery from Text is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP). The aim is to get insights into big quantities of text data. The Knowledge Discovery in Text, while deeply rooted in NLP draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management and others for its discovery process. In this research work we have proposed a classification model using Kernel based EM algorithm with Gini-Index with clustering mechanism for extracting the quantitative assessment of the efficacy drug ingredients are analyzed based on reviews with text mining method. And also it turned out that the same drugs have a difference in the effect. The experimental methods discussed are directly applicable to certain computational aspects of drug discovery. Our future work includes extracting the drugs information through registered users for effective classification of exact data in vast databases.

REFERENCES

1. A Lew and H. Mauch, "Intoduction to Data Mining Principle" ,SCI, springer, 2006.
2. Ananiadou, S. and Mcnaught, J. "Text Mining for Biology And Biomedicine", Artech House, Inc., Norwood, MA, USA, 2005.
3. Cristianini, N. and Shawe-Taylor, J. "An Introduction to Support Vector Machines(and Other Kern el-based Learning Methods)". Cambridge University Press,2000.
4. Ghahramani, Z. and Hinton, G. "The EM Algorithm for Mixtures of Factor Analyzers". Technical report, University of Toronto, 1997.
5. Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. kernlab, "An S4 package for kernel methods in R". Journal of Statistical Software, 11(9),2004.
6. Krallinger, M., Erhardt, R. A.-A., and Valencia, A. "Text-mining approaches in molecular biology and biomedicine". Drug Discovery Today, 10(6),2005.
7. N. Cancedda, N. Cesa-Bianchi, A. "Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov/pubs/trec11/papers/kernit.ps.gz, 2002.
8. Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transact ions on Knowledge and Data Engineering. 2010.

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. 3, Issue 12, December 2015

9. Nidhi Mishra et al, "Classification of Opinion Mining Techniques", International Journal of Computer Applications, Vol 56, No 13, Pg No 1-6, Oct 2012.
10. Raisa Varghese, Jayasree, "A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology, Vol 2 Issue 11 Nov 2013.
11. V. Gupta, G.S. Lehal, "A Survey of Text Mining Techniques and applications", Journal of Emerging Technologies in Web Intelligence, 2009.
12. Vishal Gupt a, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", journal of emerging technologies in web intelligence, vol. 1, no. 1, august 2009.
13. Y. Li, W. Yang, and Y. Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules", Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 953-958, 2006.
14. Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs", IEEE Trans. Mining (ICDM '03), pp. 593-596, 2003.
15. Yoon B, Park Y. "A text-mining-based patent network: analytical tool for high technology trend", J High Technol Manag Res, 15:37e50 ,2004.

BIOGRAPHY

Jyotsna Pulipati, has completed Master Degree from Sri Venkateswara University, Pursuing Ph.D in the department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India. She was Published research papers in various international journals. Her areas of interests include Data Mining, Text Mining, Image Processing, Software Engineering and Data Base Management System etc.

Dr. P. Govindarajulu, worked as a professor in the department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India. He played various roles in the department as well as in the university and published several research papers in various national, International journals, conferences and workshops. His research areas of interests include Image Processing, Data Mining, Cloud Computing, Web Mining, Data Base Management System, etc.