IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Feature Extraction and Classification of Web Data

Hemlata Patel

(Ph.D Schcolar)

Computer Science and Engineering
Department
Dr. A.P.J. Abdul Kalam University,
Indore (M.P.) - India

Dr. Dhanraj Verma
Computer Science and Engineering
Department
Dr. A.P.J. Abdul Kalam University,
Indore (M.P.) - India

Abstract— **for the last few years, text mining has been evolving and gathering revealing importance. The number of text documents in digital form is increasing and available to users through variety of sources like e-media, digital media and many more. Due to vast availability of text, a lot of unstructured data has been collected and converted into defined structured data. This process is known as text classification. High dimensionality of feature space is one of the problems in text classification. This is solved by feature selection and feature extraction methods and improves the performance of text classification. The feature extraction techniques remove the irrelevant and useless features from the text documents and reduce the dimensionality of feature space. This paper proposed a system for feature extraction and classification of text data. First features of text are extracted and then classified by classifier. The proposed solution is based on semi supervised learning. Datasets used for training and testing will be obtained from user feedback from different web sites. The results show that the proposed feature extraction and classification approach is simple, computationally tractable, and achieves low error rates.**

Keywords—text mining, text classification, feature extraction.

I. INTRODUCTION

Now a day's information technology and the Internet applications, the technology innovations, such as big data and high dimensional data have exported from IT industry to others due to rapid development. The big data of high volume which cannot be collected traditionally are getting into use. As a result, useful information is often unnoticed, and the possible benefits of increased computational and data collection capabilities are only sometimes executed. The data can be cut, managed, processed, and organized for the different purposes like medical decisions, business decisions and so on. In this process, data mining is the most important, which can take and extract more important data from different fields such as business organizations as well as social system and organizations. Text mining came into being.

Text mining is the process of discovering useful and interesting knowledge from unstructured text. In order to discover knowledge from unstructured text data, the first step is to convert text data into a manageable representation. A common practice is to model text in a document as a set of word features, i.e., "bag of words" (BOW). Often, some feature selection techniques are applied, such as stop-word removal or stemming, to only keep meaningful features and improve the accuracy using supervised classification algorithm.

We analyze real-life classification problems with high dimensional features. The results show that the proposed classification and feature selection approach is simple, computationally tractable, and achieves low error rates which are key for the construction of advanced decision support systems.

II. LITERAURE SURVEY

1. Bissan Ghaddar et al perform data mining on two types of data. They implemented a system for feature selection and classification using SVM.

2 Raj Kumar et al implemented classification algorithms for different types of data sets like data of patients, financial data according to performances.

6 S.Deepajothi1 et al. presented paper, in which they used different classification algorithms and compare their classification accuracy.

7 Manish Kumar Shrivastava et al presented the importance of data mining for different purpose different business domains.

8 M.sujatha et al studie basic concept of different feature selection methods. They reviewed four filter based feature ranking techniques and one wrapper based feature ranking technique.

10 Jundong Li jundongl et al. shows the effectiveness of feature selection in preprocessing data and reducing data dimensionality which is essential to successful data mining and machine learning applications.

11 Ms. Shweta Srivastava1 et al. provides a comprehensive overview of various characteristic of feature selection. Therefore, more efficient search strategies and evaluation criteria are needed for feature selection with large dimensionality.

12 Xia Huosong et al. Evaluated the feature selection methods in dimensionality reduction for text categorization at all the reduction levels of aggressiveness, form using the full vocabulary as the feature space, to removing 98%of the unique terms.

14 S.Vanaja K et al. Studied and survey about feature selection algorithms shows that the feature selection algorithm consistently improves the accuracy of the classifier. Each feature selection methodology has its own advantages and disadvantages. The dataset with larger attributes use the wrapper methods with lesser improvement in accuracy. Each algorithm has different behaviour which shows relaying single algorithm for different dataset is infeasible. The feature selection algorithms are one which decides the accuracy of the classification of different datasets. The feature selection algorithm must select the relevant features and also remove the irrelevant and inconsistent features which cause the degradation of accuracy of the classification algorithms.

16 Durgesh K. Srivastava et al. presented that the classification is one of the most important tasks for different application such as text categorization, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data Classification etc.

17 Zhu Jin et al. addressed nonlinear separable problem in pattern recognition, they also told that appropriate kernel function is crucial to performance of a classifier, thus, the determination of proper kernel function and its parameters is especially important to SVM.

18 Hyeran Byun et al. they have presented a brief introduction on SVMs and several applications of SVMs in pattern recognition problems. They discussed that SVMs can applied to a number of applications ranging from face detection and recognition, object detection and recognition, and so on.

22 Prof. Stéphane Canu gives a technical report on kernel methods. They teaches that kernel methods are a class of learning machine that has become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection.

28 Jigna Ashish Patel done survey on comparison among data mining classification's algorithms and analyzing of the time complexity of the mentioned algorithms.

29 Babu C Lakshmanan et al. give the introduction of data mining and its application area. They also proposed a methodology for the programmed exposure and classification to evaluate the pattern on effectiveness of treatment for Pulmonary Tuberculosis (PTB) patients.

III PROPOSED METHODOLOGY

Tools and Technology Used: Data mining tools and techniques are now more important than ever for all businesses, big or small. The development and application of data mining algorithms requires the use of powerful software tools. Now a day's many readymade tools are available. Some of them are-

a) NLTK provides a pool of language processing tools including data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks. We only need to do is install NLTK, pull a package for our favourite task and we are ready to go. We can build applications on top if it and customizing it for small tasks.

b) WEKA is a JAVA based customization tool which is used to visualization, classification and predictive analysis.

c) MATLAB is an interactive software system for numerical computations and graphics. As the name suggests, Mat lab is especially designed for matrix computations: solving systems of linear equations, computing Eigen values and eigenvectors, factoring matrices, and so forth. In addition, it has a variety of graphical capabilities, and can be extended through programs written in its own programming language. Many such programs come with the system; a number of these extend Mat lab's capabilities to nonlinear problems, such as the solution of initial value problems for ordinary differential equations.

Another most important task in data mining is to select the correct data mining technique. Data mining technique has to be chosen based on the type of business and the type of problem that we faces. There are basically seven main data mining techniques. In which cclassification is the most commonly used, which contains a set of pre classified samples to create a model which can classify the large set of data. This technique helps in deriving important information about data and metadata. There are two main processes involved in this technique.

i. Learning – In this process the data are analyzed by classification algorithm
ii. Classification – In this process the data is used to measure the precision of the classification rules

Methodology: - The main aim of this research is to investigate and develop a methodology to facilitate semi supervised learning based approach. This research is conducted to test different types of data. The proposed system is an improvement to the technique introduced in noteworthy contributions. Fig. 1 shows the general structure of the system.

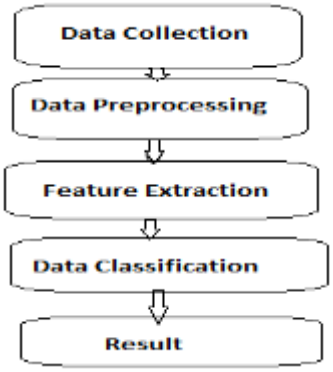

Figure 1: General Structure of System

a) *Data Collection:* - Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion. For the research purpose a lot of data is available on UCI Repository. These are used to maintain data as a service and contain lots of datasets. We may view and search any type of data sets through their searchable interface.

b) *Data Preprocessing:-* Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user -- for example, There are a number of different tools and methods used for preprocessing, including: sampling, transformation, demising, normalization,.

c) *Feature Extraction:* - Feature extraction is the process in which certain raw features are extracted and used in decision making for further processing.

d) *Data Classification:-* Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various business or personal objectives. For example separating customer data based on gender, Identifying and keeping frequently used data in disk/memory cache, Data sorting based on content/file type, size and time of data

e) *Result:* - it is a final step that shows the result of testing data.

In the proposed system Semi-Supervised Feature Selection method is used. It is a combination of unsupervised and supervised learning methods. However, in many real-world applications, we often have a small number of labeled samples and a large number of unlabeled samples. Both supervised and unsupervised feature selection algorithms are not suitable in this scenario. For supervised methods, the small number of labeled samples may be insufficient to provide correlation information of features; while unsupervised methods totally ignore class labels which could provide useful information to discriminate different classes. Therefore, it is desirable to develop semi-supervised methods by exploiting both labeled and unlabeled supervised feature selection Figure 2 shows the general structure of the system.

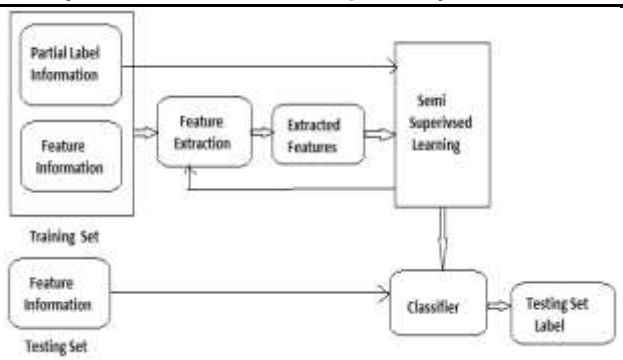

Figure 2: Proposed System

There are two modules of proposed system. One for the unsupervised learning and other for the supervised learning.

6. Expected outcome of the proposed work

The expected outcome of this research is an improved understanding of the operation of the system. The benefits of this improved understanding will principally be felt in improved private and public decision making and improved performance of the system. It include building simpler and more comprehensible models, improving data mining performance, and helping prepare, clean, and understand data. It is also shown that data mining technology can be used in many areas in real life including medical, financial, the retail industry and also in the social networking. One of the biggest challenges for data mining technology is managing the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. Having optimized data mining and machine learning techniques it helps the decision makers to take precise decision about the organization to gain more incremental profit.

REFERENCES

[1] Bissan Ghaddar, Joe Naoum-Sawaya, (2017), High Dimensional Data Classification and Feature Selection using Support Vector Machines, *European Journal of Operational Research*

[2] Raj Kumar, Dr. Rajesh Verma, (2012), Classification Algorithms for Data Mining: A Survey, IJIET Vol. 1 Issue 2 ISSN: 2319 – 1058

[3] S. Deepajothi1, Dr. S. Selvarajan, (2012), A Comparative Study of Classification Techniques on Adult Data Set, IJERT Vol. 1 Issue 8 ISSN: 2278-0181

[4] Manish Kumar Shrivastava, Praveen Chouksey, Rohit Miri, (2013), Exploring Data Mining Classification Techniques, IJERT Vol. 2 Issue 6 ISSN: 2278-0181.

[5] M. Sujatha, Dr. G. Lavanya Devi, (2013), Feature Selection Techniques using for High Dimensional Data in Machine Learning, IJERT Vol. 2 Issue 9 ISSN: 2278-0181.

[6] Jundong Li jundongl, Kewei Cheng,Suhang Wang,Fred Morstatter ,Robert P. Trevino ,Jiliang Tang , Huan Liu , Feature Selection: A Data Perspective

[7] Ms. Shweta Srivastava, Ms. Nikita Joshi, Ms. Madhvi Gaur, (2013), A Review Paper on Feature Selection Methodologies and Their Applications, International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, Volume 7, Issue 6, PP. 57-61.

[8] Xia Huosong, Liu Jian, (2011), The Research of Feature Selection of Text Classification Based On Integrated, 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 978-0-7695-4415-1/11.

[9] S. Vanaja, K. Ramesh kumar, (2014), Analysis of Feature Selection Algorithms on Classification: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 96– No.17.

[10] Durgesh K. Srivastava, lekha Bhambhu, (2009), Data Classification Using Support Vector Machine, Journal of Theoretical and Applied Information Technology © 2005 - 2009 Jatit.

[11] Zhu Jin, Xiaoping Ma, (2010), Model Selection for Support Vector Machines Based on Kernel Density Estimation, 978-1-4244-5182-1/10/ IEEE.

[12] Hyeran Byun, Seong-Whan Lee, (2002), Applications of Support Vector Machines for Pattern Recognition: A Survey, Springer-Verlag Berlin Heidelberg, LNCS 2388, pp. 213-236.

[13] Prof. Stéphane Canu, (2014), SVM and kernel machines: linear and non-linear classification, Ocean's Big Data Mining.

[14] Jigna Ashish Patel, (2015), Classification Algorithms and Comparison in Data Mining, International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4.

[15] Babu C Lakshmanan, Valarmathi Srinivasan, Chinnaiyan Ponnuraja, (2015), Data Mining with Decision Tree to Evaluate the Pattern on Effectiveness of Treatment for Pulmonary Tuberculosis: A Clustering and Classification Techniques, SCIRJ, Volume III, Issue VI, ISSN 2201-2796