Categorize Online news Using Various Classification Techniques

Neeru Sharma, Paramjit Kaur

ABSTRACT- Classification is a data mining technique used to predict group membership for data instances. It is often referred as "supervised learning". It has a predefined set of groups or models based on that we predict value [2]. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Since the emergence of WWW, it is essential to handle a very large amount of electronic data of which the majority is in the form of text. This scenario can be effectively handled by various Data Mining techniques. This paper proposes an intelligent system for classify the inner structures of the online news based on Neural Network (NN) and Support Vector Machine (SVM).For the current scenario, the work has just been done to identify the outer clusters of the system but no work till now has been done for inner cluster of the datasets. In this proposed work, we would be creating inners clusters for each and every field of the proposed system like Sports, Entertainment and Financial. In this work we would be creating clusters for Sports, Entertainment and Financial also so that we can go on for better accuracy.

Index Terms—Text Classification, Support Vector Machines (SVM), Neural Network, Online News.

INTRODUCTION:

Data mining is a powerful new technology to help companies focus on most important information in the data. It is the process of analyzing data from different perspectives and summarizing it into useful information [5]. Mining the data means fetching out a piece of data from a huge data block. Nowadays corporate and organizations are accumulating data at an enormous rate and from a very broad variety of sources such as customer transactions, credit card transactions, bank cash withdrawal to hourly weather data. To put the data into the database servers, online transactional process (OLTP) systems have been developed to help business run smoothly based on their own business processes. Those OLTP systems stores all the transactional data into the database for every transaction happens to the business in every second such as sale orders, purchase orders in sale to head count data in human capital management[12]. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought.

Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Online news classification has been challenge always in terms of manual transaction. News contents are one of the most important factors that have influence on market. Easy and quick availability to news information was not possible until the beginning of the last decade. The goal of classification is to accurately predict the target class for each case in the data. For example model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. Classification are discrete and do not imply order. Classification has many applications in customer segmentation, business modeling, marketing, credit analysis and drug response modeling. Classification is always based on two things:

a) The area which you choose for the classification that is the cluster region.

b) The kind of dataset which you are going to apply on the selected region [1].

News articles on topical issues are helpful for company managers and other decision-makers. However, due to the sheer number of news articles published, it is a time-consuming task to select the most interesting one. Therefore, a method of news-article categorization is essential to obtain the relevant information quickly. The large amount of news being generated these days through various websites, it is possible to mine the general sentiment of a particular company being portrayed by media agencies over a period of time, which can be utilized to gauge the long term impact on the investment potential

ISSN: 2278 – 1323

of the company. Nowadays, there is a large amount of information available in the form of text in diverse environments, the analysis of which can provide many benefits in several areas. Steps involved in news classification:

- News Pre-processing
- Organization detection
- Keyword Detection
- Headline Preprocessing
- Detection of Products
- Executives Detection
- Feature Generation

Classification Techniques:

Decision Trees: A Decision Tree text classifier is a tree in which internal nodes are labeled by terms, branches departing from them are labeled by the weight that the term has in the text document and leafs are labeled by categories [2]. Decision Tree constructs using 'divide and conquer' strategy. Each node in a tree is associated with set of cases. This strategy checks whether all the training examples have the same label and if not then select a term partitioning from the pooled classes of documents that have same values for term and place each such class in a separate sub tree [8].

Naïve Bayes Algorithm: Naïve Bayes classifier is a simple probabilistic classifier based on applying Baye's Theorem with strong independence assumptions. This algorithm computes the posterior probability of the document belongs to different classes and it assigns document to the class with the highest posterior probability [8].

Rocchio's Algorithm: Rocchio's learning algorithm is in the classical IR tradition. It was originally designed to use relevance feedback in querying full-text databases, Rocchio's Algorithm is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class [7] and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

K. Nearest Neighbor: KNN is a classification algorithm as given in where objects are classified by voting several labeled training examples with their smallest distance from each object. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification [8]. This method performs well even in handling the classification tasks with multi-categorized documents.

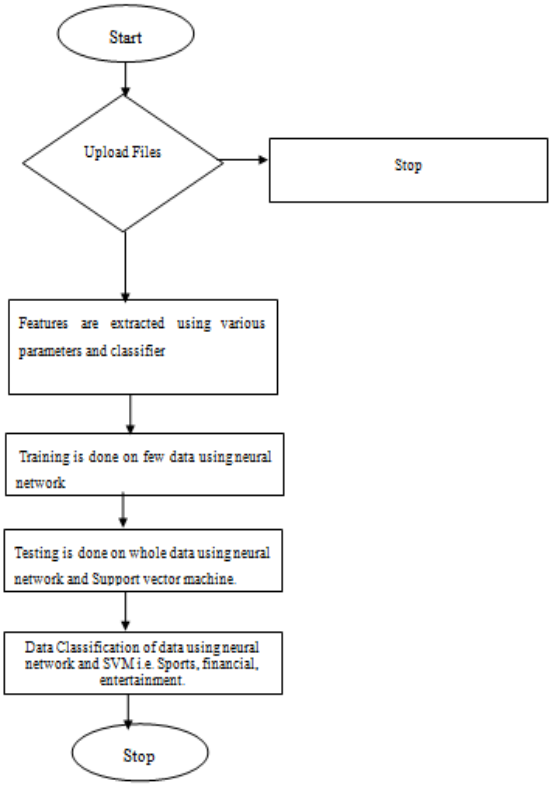
Its disadvantage is that KNN requires more time for classifying objects when a large number of training examples are given. KNN should select some of them by computing the distance of each test objects with all of the training examples.

Neural Network: A neural network classifier is a network of units, where the input units usually represent terms, the output unit(s) represents the category. For classifying a test document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit(s) takes up as a consequence determines the categorization decision [5]. Some of the researches use the single-layer, due to its simplicity of implementing. The multi-layer which is more sophisticated, also widely implemented for classification tasks.

Support Vector Machines (SVM): A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task. The SVM need both positive and negative training set which are uncommon for other classification methods [7]. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector.

PROPOSED METHODOLOGY:

The proposed news classifier is designed and developed for inner classification of news using SVM (Support Vector Machine) and Neural Network. Flow chart of proposed system is as shown in figure.



ISSN: 2278 – 1323

Fig 1: Proposed Flow Chart

EXPERIMENTAL RESULTS:

To evaluate the news classification method proposed in this paper, we collect the news from various newspapers related to financial, sports, and entertainment in Table 1.

CATEGORIES	SUBCATEGORIES	NEWS PROCESSED
ENTERTAINMENT	BOLLYWOOD SERIALS	30 28
FINANCIAL	FUNDS ACCOUNTS	35 30
SPORTS	CRICKET FOOTBALL	32 25

Table 1. Processed News of Different Categories

Total input news related with boll wood is 30 but it classify 27 news correctly. Total Serial news as input is 28 and it classify 22 correctly. Among the 35 funds news it classify 31 news correctly. Input news of Accounts is 30 but 25 classify correctly. In Cricket total news as input is 32 and 29 correctly, in football total news as input 30 and 25 correctly. Table 2 shows the percentage of accuracy for different news.

CATEGORIES	SUBCATEGORIES	ACCURACY
ENTERTAINMENT	BOLLYWOOD SERIALS	90.09% 89.02%
FINANCIAL	FUNDS ACCOUNTS	91.57% 93.00%
SPORTS	CRICKET FOOTBALL	93.88% 92.05%

Table.2 Percentage of Accuracy of Different News.

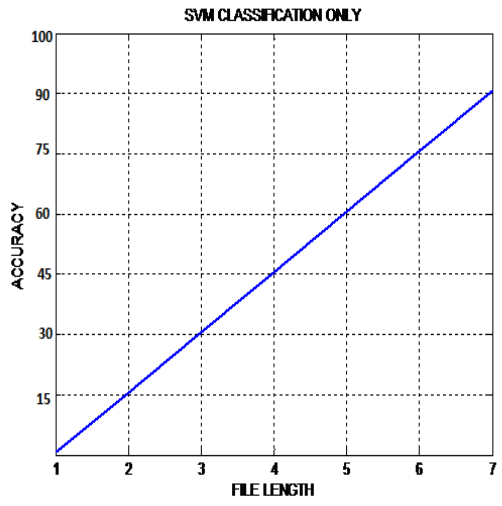


Fig 2. shows the accuracy graph using SVM classification

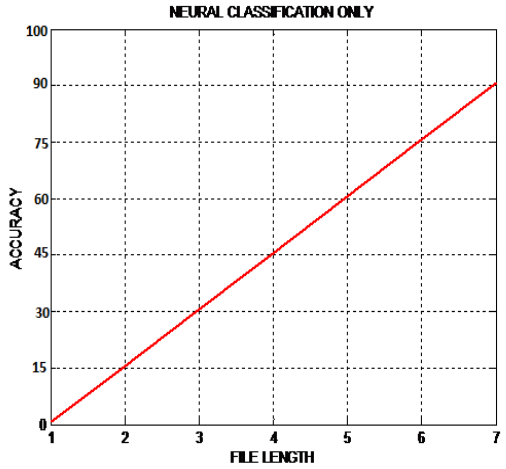


Fig. 3 shows the accuracy graph using Neural Classification

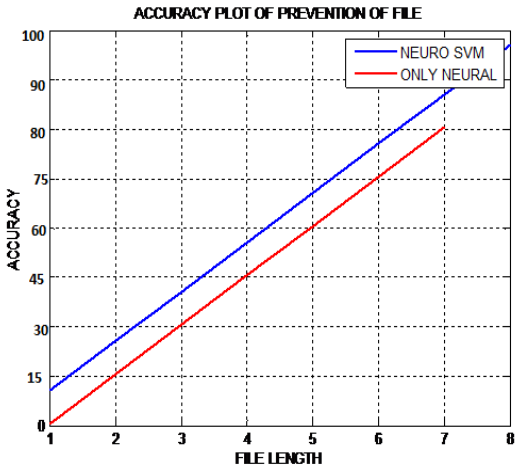


Fig 4. shows the accuracy comparison on the basis of accuracy between neural and SVM.

CONCLUSION AND FUTURE SCOPE:

In the work till now, a successful implementation has been done to extract the news from the online portals for the further processing. In addition to it the clusters of different categories has been also created so that the combination of Neural Network and Support Vector Machines (SVM) could be applied to it to regain a better efficiency.

In future the work can be made to the inner clusters of the made clusters till now.ID3 or Greedy algorithm can also be applied instead of Neural Network and SVM algorithm to check that whether the applied algorithm has better implementation result or not.

REFERENCES:

[1]. Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad " Financial News Classification using SVM"International Journal of

ISSN: 2278 – 1323

Scientific and Research Publications, Volume 2, Issue 3, March 2012 .

[2]Vandana Korde and C Namrata Mahender "Text Classification and Classifiers:A Survey" International [3]. Mofleh Al-diabat "Arabic Text Categorization Using Classification Rule Mining "Applied Mathematical Sciences, Vol. 6, 2012, no. 81, 4033 - 4046 "

[4].Koby Crammer, Mark Dredze and Fernando Pereira" Confidence-Weighted Linear Classification for Text Categorization"" Journal of Machine Learning Research 13 (2012) 1891-1926"

[5] M. Pushpa, Dr. K. Nirmala " Text Categorization Using Activation Based Term Set" International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[6] Megha Dawar and Dr. Aruna Tiwari "Fast Fuzzy Feature Clustering for text Classification" International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, Janvery 2012.

[7] Durga Bhavani Dasari & Dr.Venu Gopala Rao. K "Text Categorization and Machine Learning Methods: Current State of the Art" Global Journal of Computer Science and Technology Software & Data Engineering Volume 12 Issue 11 Version 1.0 Year 2012.

[8] Y. Zhai, A. Hsu, and S. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," Lecture Notes in Computer Science, 2007, pp. 1087-1096.

[9] Pratiksha Y. Pawar and S. H. Gawande"A Comparative Study on Different Types of Approaches to Text Categorization" International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[10] Cingiz, M.O" Content Mining of Microblogs" 2012 IEEE International Conference on Advances in Social Networks Analysis and Mining 26-29 Aug. 2012 Page(s): 835- 838

[11] Samarawickrama, Sameendra" Effect of Named Entities in Web Page Classification" 2012 IEEE Fourth International Conference on Computational Intelligence, Modelling and Simulation Sept. 2012 Page(s): 38- 42

[12] Wang, Jenq-Haur "Statistical Single-Document Summarization for Chinese News Articles" 2012 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA).

Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[13] Syed Aqueel Haider and Rishabh Mehrotra" Corporate News Classification and Valence Prediction: A Supervised Approach"International Journal on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 175–181,24 June, 2011.

[14]Thair Nu Phyu" Survey of Classification Techniques in Data Mining" International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009.

[15] Wu, Chung-Hsien H." Story Segmentation and Topic Classification of Broadcast News via a Topic-Based Segmental Model and a Genetic Algorithm" IEEE Transactions on Audio, Speech, and Language Processing Nov. 2009.

[16] Roshan Sumbaly,Shakti Sinha"Sentiment Mining in Large News Datasets.",May10, 2009.

[17] James c.Bezdek, Robert Ehrlich and William Full "FCM: The Fuzzy c-means clustering algorithm" Computers & Geosciences vol.10, no.2-3, pp 191-203, 1984

[18] Li W. T., Shi X. W., Xu L. and Hei Y.Q. , "Improved GA and PSO Culled Hybrid Algorithm for Antenna Array Pattern Synthesis", Progress In Electromagnetics Research, PIER 80, (2008), pp. 461–476.

[19] Shi, X.H., Liang Y.C., Lee H.P., Lu C. and Wang L.M., "An Improved GA and a Novel PSO-GA-Based Hybrid Algorithm", Information Processing Letters, Vol. 93, No. 5, (2005), pp. 255-261

[20] Tang J, Zhang G, Lin B and Zhang B, "A Hybrid PSO/GA Algorithm for Job Shop Scheduling Problem", Advances in Swarm Intelligence, Springer Berlin, Vol. 6145, (2010), pp. 566-573.

Neeru Sharma is pursing Master Degree in Computer Science from Indo Global College of engineering, Abhipur , Mohali, Punjab .Her Present research work in Categorize the online news using classification techniques.

Paramjit Kaur Assistant Professor in Computer Science in Indo Global College of engineering, Abhipur , Mohali, Punjab. Her Present research work in Categorize the online news using classification techniques.

ISSN: 2278 – 1323