INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume2, Issue6)

Available online at: www.Ijariit.com

K-Means Clustering based Lexicon Analytical Model for Multi-Source News Classification

Kamaldeep Kaur[1], Maninder Kaur[2]
Doaba Institute of Engineering & Technology, Kharar
deepsandhu9190@gmail.com , maninderecediet@gmail.com

Abstract: The supervised models have been found more efficient for the purpose of news classification. The major goal of the news classification research is to improve the accuracy while decreasing the elapsed time. It is always difficult for the people to read all of the news on their favourite's portal which have listed over the given portal. In this research, an approach is KNN lexicon technique which is used to obtain the popular news list from thousands or hundreds of online news available through APIs. This approach uses extraction summarization for summarizing the keywords thereby selecting the original sentences and putting it together into a new shorter text explaining the overall overview of the news data. Then the lexicon analysis would be performed over the given text data and then final classification of the news is done using k-nearest neighbor. The results would be obtained in the form of the parameters of accuracy, elapsed time, etc.

Keywords: News Classification, Regression, Probabilistic Classifier, Automatic Categorization, Multi-domain news analysis.

I. INTRODUCTION

In general, new classification tasks may be classified into 2 categories: descriptive data processing and prophetical data processing. The previous describes the info set in laconic outline manner and presents attention-grabbing general properties of information. An information mining system might accomplish one or a lot of the subsequent data processing tasks.

- Category Description: Class description provides a laconic and account of assortment of information and distinguish it from others. The account of assortment of information is named category characterization, the comparison between 2 or a lot of collections of information is named comparison or discrimination.
- Association: Association is discovery of association relationships or correlations among a group of things. There are varied association analysis algorithms like Apriori search, mining multiple level, multi dimensional association, mining association for numerical.
- Classification: Classification analyzes a group of coaching information (a set of object whose category label is known) and constructs a model for every category supported the options within the information. A choice tree or set of classification rule is

generated by such a classification method. There are several classification technique developed within the field of machine learning, static, database, neural network.

- Prediction: This mining performs predicts the attainable price of some missing information and also the price distribution of certain attributes during a set of objects. It involve the finding of set of attributes relevant of the attribute of interest and predicting the worth distribution supported set of information almost like choose object.
- Agglomeration: Clustering analysis is tool established clusters embedded in information wherever a cluster may be a assortment of information object that's almost like each other. Similarity may be such that by user of consultants.

Time Series Analysis: statistic analysis is to research giant set of your time series information to seek out bound regularities and attention-grabbing characteristics, together with rummage around for similar sequences and sub sequences, mining consecutive patterns, regularity, trends and deviation.

II. LITERATURE REVIEW

Ouyang, Yuanxin et. Al., 2014, [22] has projected the news title classification with support from auxiliary long texts. During this paper, the authors have targeted on the matter of reports title classification that is a vital and typical member in brief text family and propose an approach that employs external info from long text to deal with the matter the scantiness.

Prollochs, Nicolas et. Al., 2014, [9] has worked on the sweetening of sentiment analysis of monetary news by detective work negation scopes. To predict the corresponding negation connected and related literature content, which is usually, utilized on the basis of the hybrid approaches, which are called the rule-based algorithms and the expert learning based machine learning mechanism.

V Bolon-canedo et al., 2014, [5] proposes information classification practice an ensemble of filters throughout this analysis, the thought of assembling is ready-made for feature selection. Authors propose an ensemble of filters for classification, double-geared toward achieving a good classification performance at the aspect of a reduction among the input property.

Li, Jinyan et. Al., 2015, [17] have projected the hierarchic classification in text mining for sentiment analysis of on-line news. During this paper, the authors have evaluated many widespread classification algorithms, at the side of 3 filtering schemes.

Ronny luss et al., 2015 [19] proposes Predicting Abnormal Returns from News mistreatment Text Classification. They show however text from news articles is accustomed predict intraday worth movements of monetary assets mistreatment support vector machines.

III. EXPERIMENTAL DESIGN

The ranking algorithm has been used to fetch the words out of the news data, which are further used to evaluate the class of the news data. The word list or keyword list contains the ranking data for different news categories of politics, business, sports, entertainment and technology. The ranking algorithm matches the news data will all keywords lists one by one and gives the rank values in the Ranking vector on the position where the word in the news data matches the word in the keyword list.

Algorithm 1: Ranking Index algorithm

```
1. Acquire the news data from the online source or local source
2. Extract the ontology method based keywords from the given news text
3. Apply the keyword matching and weight calculation using the supervised method with the specific category based list matching method
4. Construct the keyword matching matrix using the pre-defined weight lists stored int eh SRD (Sparse Ranking  Data).
5. Iterate the steps 3 and 4 iteratively for all news texts
```

In the Ranking algorithm, all training samples were used for training, that is, whenever the news classification sample or test sample needs to be verified, analyzed and classified, it is necessary to calculate similarities between that sample and all documents in the training sets, and then choose Ranking with word samples which have largest similarities.

At first, perform the weight computational algorithm over the keyword data after the successful application of the pre-processing based method to assess the real-time term weight and frequency.

Algorithm 2: Weighted value based k-means with random point clustering algorithm

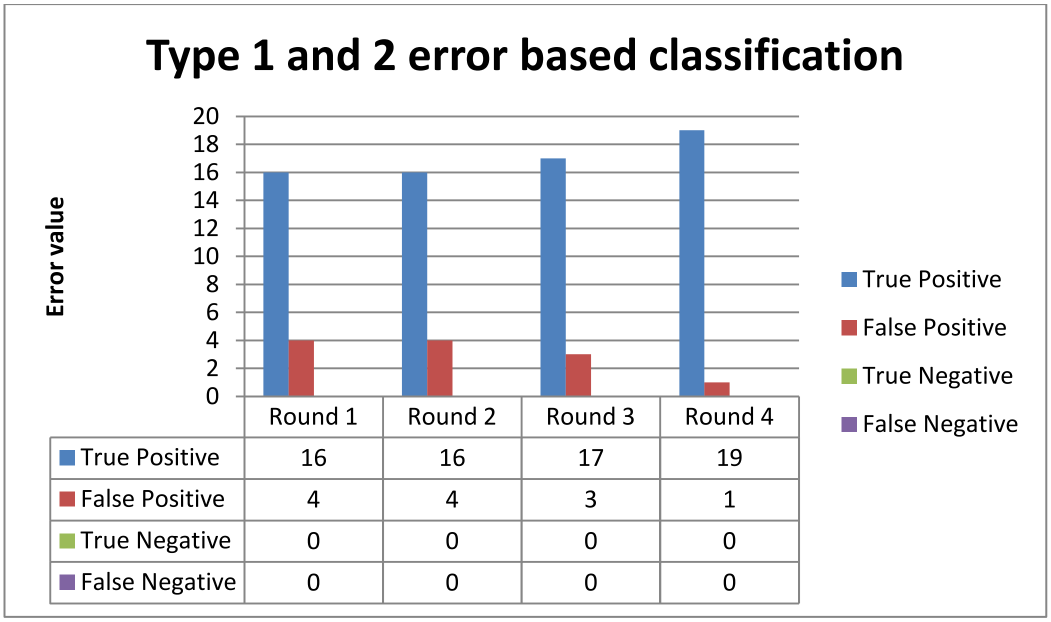
1. Initialize the value of the clusters, denoted K, for the segmentation of the given document.
2. Obtain the random position for centroid equals the number K, from the pre-defined set of the centroid
3. Find the distance of the object or keyword from the selected centroid.
4. Perform the object assignment to the nearest centroid or with the minimum distance
5. Update the centroid value, if the condition floats below the average satisfaction value.
6. Iteration from the step 3 to 5 for every keyword or object to classify the whole data.

IV. RESULT ANALYSIS

Table 1: The statistical errors obtained from the simulation

Error Type	Round 1	Round 2	Round 3	Round 4
True Positive	16	16	17	19
False Positive	4	4	3	1
True Negative	0	0	0	0
False Negative	0	0	0	0

The table 1 contains the statistical parameters in account from the experiments conducted over the KNN lexicon technique. The KNN lexicon technique has been obtained with the primary statistical type 1 and type 2 errors. In the final, the accuracy has been increased due to the higher convergence.



	Round 1	Round 2	Round 3	Round 4
🟦 True Positive	16	16	17	19
🟥 False Positive	4	4	3	1
🟩 True Negative	0	0	0	0
🟪 False Negative	0	0	0	0

Figure 1: Type 1 and 2 errors recorded from the KNN lexicon technique simulation

The figure 1 is visualizing the statistical parameters obtained under the table 1. The clearly visible results of round 4 clearly justify the robustness of the KNN lexicon technique for the higher order accuracy for the automatic news classification engine.

CONCLUSION

A number of experiments have been conducted over KNN lexicon technique by using the various forms of the input data generated after various levels of pre-processing. The proposed technique has been tested for the various performance measures which includes the precision, recall, average prediction accuracy and F1-measures. All of the above performance measures has been obtained after the estimation of the statistical type 1 and type 2 errors over the input data. The proposed technique has been found accurate higher than 90% in all of the rounds if the true negative cases are also being analyzed. The proposed technique has been recorded with the average accuracy over all of the test cases nearly at 93% which is better all of the other models used under the existing model. The KNN lexicon technique has outperformed all of the existing models designed with the different filters over the differently processed datasets.

REFERENCES

[1] Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining in education: data classification and decision tree approach." International Journal of e-Education, e-Business, e-Management and e-Learning 2, no. 2 (2012): 140.

[2] Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. "Sentiment analysis in the news." arXiv preprint arXiv:1309.6202 (2013).

[3] Bhowmick, Plaban Kumar, Anupam Basu, and Pabitra Mitra. "Classifying emotion in news sentences: When machine classification meets human classification." International Journal on Computer Science and Engineering2, no. 1 (2010): 98-108.

[4] Bhukya, Devi Prasad, and S. Ramachandram. "Decision tree induction: an approach for data classification using AVL-tree." International Journal of Computer and Electrical Engineering 2, no. 4 (2010): 660.

[5] Bolón-Canedo, Verónica, Noelia Sánchez-Marono, and Amparo Alonso-Betanzos. "Data classification using an ensemble of filters." Neurocomputing135 (2014): 13-20.

[6] Byun, Hyeran, and Seong-Whan Lee. "Applications of support vector machines for pattern recognition: A survey." In *Pattern recognition with support vector machines*, pp. 213-236. Springer Berlin Heidelberg, 2002.

[7] CREŢULESCU, Radu George, and N. VINŢAN Lucian. "Contributions to Document Classification System Design." Vol. 1, Issue 1, SIBIU, 2011.

[8] Cui, Limeng, Fan Meng, Yong Shi, Minqiang Li, and An Liu. "A Hierarchy Method Based on LDA and SVM for News Classification." In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pp. 60-64. IEEE, 2014.

[9] Denial I.Morariu, Lucian N. Vintan, and Volker Tresp, "Meta- Clalssification using SVM classifiers for text documents ", "World Academy of science engineering and technology" 21,2006.

[10] D. Morariu, R. Cre,Tulescu and L.,Vin,tan, " improving the SVM Meta Clssifier for text document by using Naïve bayes,"Int. J. of Computers, communication and control, ISSN 1841-9844.

[11] Durgesh, K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." Journal of Theoretical and Applied Information Technology12, no. 1 (2010): 1-7.

[12] Ge, Jiaqi, Yuni Xia, and Chandima Nadungodage. "UNN: a neural network for uncertain data classification." In Advances in Knowledge Discovery and Data Mining, pp. 449-460. Springer Berlin Heidelberg, 2010.

[13] Hyeran Byun 1 and Seong-Whan Lee2, " Application of Support Vector machines for pattern recognition: A Survey," SVM 2002, LNCS 2388,pp.213-236,2002.

[14] Kirange, D. K. "Emotion classification of news headlines using SVM." Asian Journal of Computer Science & Information Technology 2, no. 5 (2013).

[15] Krishnlal G, S Babu Rengarajan, K G Srinivasagan, " A new text mining approach based on HMM-SVM for web news classification" International Journal of Computer Applications (0975-8887) Volumn 1- No.19,2010.

[16] Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3, no. 2 (2012): 85.

[17] Li, Jinyan, Simon Fong, Yan Zhuang, and Richard Khoury. "Hierarchical classification in text mining for sentiment analysis of online news." *Soft Computing* (2015): 1-10.

[18] Lie Lu, Stan Z. Li and Hong –Jiang Zhang, "Content based Audion Segmentation using Support vector machine."

[19] Luss, Ronny, and Alexandre d'Aspremont. "Predicting abnormal returns from news using text classification." Quantitative Finance 15, no. 6 (2015): 999-1012.

[20] Mita K. Dalal, Mukesh A.Zaveri," Automatic text Classification," International Journal of Computer Applications (0975-8887) Volumn 28- No.2, August 2011.

[21] Morariu, Daniel, L. Vintan, and Volker Tresp. "Feature Selection Methods for an Improved SVM Classifier." In *Proceedings of 14th International Conference on of Intelligent Systems (ICIS06), ISSN*, pp. 1305-5313. 2006.

[22] Ouyang, Yuanxin, Yao Huangfu, Hao Sheng, and Zhang Xiong. " target on the problem of news title classification which is an essential and typical member in short text family." In *Neural Information Processing*, pp. 581-588. Springer International Publishing, 2014.

[23] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6, no. 2 (2013): 256-261.

[24] Prollochs, Nicolas, Stefan Feuerriegel, and Dirk Neumann. "Enhancing Sentiment Analysis of Financial News by Detecting Negation Scopes." In*System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pp. 959-968. IEEE, 2015.