Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -2, October 2021, pp. 18-22 (5 pages)

Data Classification in Web Usage Mining using SVM, SVR and K-NN

HEMLATA PATEL[1], DHANRAJ VERMA[2]

[1 2]*Department of Computer Science, Dr. A. P. J. Abdul Kalam University, Indore 452016, India*
Corresponding Author Email: hemlatapatel55@gmail.com

Abstract— ***The internet and other communication channels are rapidly growing in recent years. This results in significant traffic and data in internet networks and web servers. The internet is a suitable, highly available and low cost publishing medium. Therefore a significant data is hosted and published using websites. In this domain some amount of data is directly present for common people and some of data is converted into useful information. The processing of raw data and obtaining fruitful results are known as data mining. Similarly the data mining algorithms when applied on web data then it is called web data mining or web mining. These paper represents suitable and efficient classifier for web content classification. Therefore the two feature selection techniques are utilized with the three popular supervised learning classifiers namely SVM, SVR and k-NN. The comparative performance study demonstrates the SVM and SVR is superior classifier as compared to k-NN.***

Index Terms— GINI Index, Information Gain, K- Nearest Neighbor, Support Vector Machine, Support vector regression, Web Data Mining.

I. INTRODUCTION

Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner [1]. A rich verity of information and data is generated each day in online sources. The manual analysis and utilization of knowledge is a complex task, thus we need new generation analytics techniques that refine and provide impure information. Such techniques can help in various different areas of applications such as education, research, and others. The web mining is a classical domain of research additionally it is beneficial for various real world applications. The web mining can be categorized in three parts web usages mining, content mining and structure mining. The web structure mining is a technique of link and web page structural analysis to enhance the end user experience. But the web content mining and usage mining is core of entire web. The content mining help to organize and optimize the contents of web pages in internet and usages mining demonstrate the behavior of user or user experience with a web application. In this presented work the web content mining and web usage mining is the main area of study. Here first web mining techniques are investigated, specifically based on content mining and web usages mining. Secondly the feature selection methods are explored and finally the classification algorithms are used to classify the content.

Therefore the improvement on existing techniques in order to enhance the productivity of the existing methods is also a considerable effort. Thus the proposed work is motivated to explore the existing applications of web data mining and their importance in real world.

II. LITERATURE REVIEW

This section reports the review of existing work and contributions for web mining techniques and their application in improvement in new generation learning methodology. Dutt et al provides over three decades long (1983–2016) systematic literature review on clustering algorithm and its applicability and usability in the context of EDM [2].Kumar et al are trying to give a brief idea regarding web mining concerned with its techniques, tools and applications [3].

F. Wu et al propose a semi-supervised multi-view individual and sharable feature learning (SMISFL) approach, which jointly learns multiple view-individual transformations and one sharable transformation to explore the view-specific property for each view and the common property across views [4].

S. Kumar et al propose a (i) Dual-Margin Multi Class Hypersphere Support Vector Machine classifier approach to automatically classifying web spam by type, (ii) introduce vel cloaking-based spam features which help classifier model to achieve high precision and recall rate [5].

A. P. G. Plaza et al introduce a fuzzy term weighing approach that makes the most of the HTML structure for document clustering [6].

H. Khalifi et al, semantic relationships or other approaches such as machine learning techniques can be applied to select the appropriate results to return [7].

The goal of M. O. Samuel et al, review is to make available a comprehensive and semi-structured overview of WCM methods, problems and solutions proffered. They have 57 publications including journals, conferences, and workshops in the period of 1999- 2018 as a review on this subject[8].

III. PROPOSED METHODOLOGY

Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -2, October 2021, pp. 18-22 (5 pages)

Supervised learning classifiers are implemented for finding superior combination of the classifier and feature selection technique for effective web content mining. In this context the required web content mining model is demonstrated in figure 1.

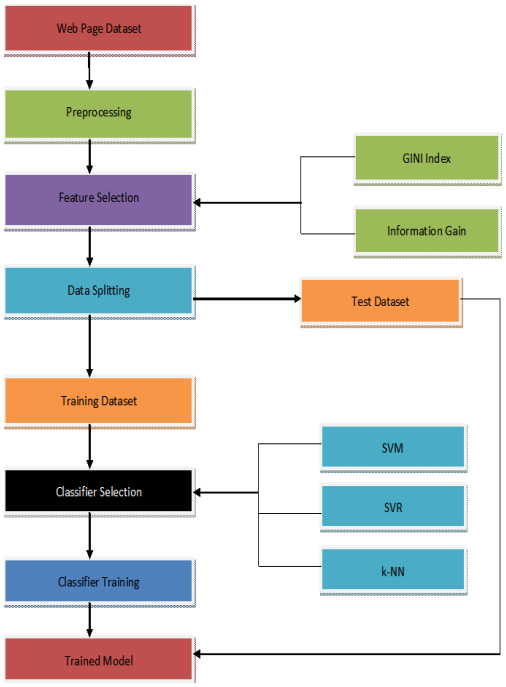

Figure 1: Classification System

1)Web Page Dataset: we had downloaded a significant amount of web pages from different subjects and designed a syntactic dataset. The data set is organized in a way by which the subdirectory consist of the class labels and the directory contents or web pages are treated as data instances to be classify in target subjects or domains.

2) Data Preprocessing: The entire web data preprocessing involve three main steps: a) Removal of HTML tags, b) Removal of special characters, and c)Removal of stop words.

3) Feature Selection: That technique helps to reduce the data dimension and regulate the requirements of the computational resources such as time and memory. In this work we involve four popular feature selection techniques used for web content mining.

a) GINI index: let S is the set of samples and having k number of classes $(c_1, c_2, \dots, c_k)$. According to the classes we define k sub categorize of data such that $\{1, 2, \dots, k\}$. Then GINI index of S can be defined using [9].

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \quad (1)$$

Where p_i is the probability which is calculated using i^{th} sample of S and complete set of S. however the minimum value of GINI is 0, which shows maximum utility of data. Similarly if the distribution of class and data is uniform then the GINI demonstrate the maximum value to 1 which shows minimum utility of data. In order to use the technique for text classification it is used as a measuring function of data impurity with respect to class labels associated with data. So according to previous consideration the lower value of GINI indicates the higher applicability of the attribute for classification.

b) Information gain: Information Gain (IG) measures how the features are. In text analysis, IG is used to measure the relevance of attribute A in class C. The higher the value of IG between classes C and attribute A, demonstrate the higher the relevance between classes C and attribute A [10].

$$I(C, A) = H(C) - H(C|A) \quad (2)$$

Where, $H(C) = -\sum_{cEC} p(C) log\, p(C)$, the entropy of the class, and $H(C|A)$ =1 is the conditional entropy of class given attribute, $H(C|A) = -\sum_{cEC} p(C|A) log\, p(C|A)$. Since Cornell movie review dataset has balanced class, the probability of class C for both positive and negative is equal to 0.5. As a result, the entropy of classes H(C) is equal to 1. Then the information gain can be formulated as:

$$I(C, A) = 1 - H(C|A) \quad (3)$$

The minimum value of $I(C, A)$ occurs if only if $H(C|A) = 1$ which means attribute A and classes C are not related at all. On the contrary, we tend to choose attribute A that mostly appears in one class C either positive or negative. On the other words, the best features are the set of attributes that only appear in one class. It means the maximum $I(C, A)$ is reached when P (A) is equal to $P(A|C_1)$ resulting in $P(C_1|A)$ and $H(C_1|A)$ being equal to 0.5. When $P(A) = P(A|C_1)$, then the value of $P(A) = P(A|C_2)$ results in $P(C_2|A) = 0$ and $P(C_2|A) = 0$. The value of $I(C, A)$ is varied from 0 to 0.5.

4) Data Splitting: After feature selection of the approach the system returns a feature vector which is used further for experimentation or learning with the supervised learning algorithm.

5) Training Set: The data splitting create two sub sets of entire web content data features first 70% of randomly selected data instances are used here for the classifier training.

6) Classifier selection: in this work the three supervised learning algorithms are implemented. In order to carried out experiments the provision is made to select an appropriate classifier according to requirements.

a) SVM: Support Vector Machine (SVM) is a supervised learning algorithm which can be used for classification or regression. It is mostly used in classification problems. In the SVM, we plot data item as a point in n-dimensional space,

where n is number of features with the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes. Support Vectors are simply the co-ordinates of observation. The SVM is a frontier which best segregates the two classes.

b) Support Vector Regression (SVR): SVR is quite different than other Regression models. It uses the Support Vector Machine to predict a variable. Other regression models try to minimize the error between the predicted and the actual value, and SVR tries to fit the best line within a predefined or threshold. It tries to classify all the prediction lines in two types, ones that pass through the error boundary and ones that don't. Those lines which do not pass the error boundary are not considered as the difference between the predicted value and the actual value has exceeded the error threshold, (epsilon). The lines that pass, are considered for a potential support vector to predict the value of an unknown.

c) **K-Nearest Neighbour (K-NN):** It is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

9) Classified Data and Performance: based on test data classification using the trained SVM the system measures the performance in terms of accuracy and error rate. At the same time the system also computes the efficiency of the system in terms of time consumed and memory usages.

IV. IMPLEMENTATION

Using the developed user interface we have tried to deliver the functional aspects of the proposed framework. The design and explanation are given as:

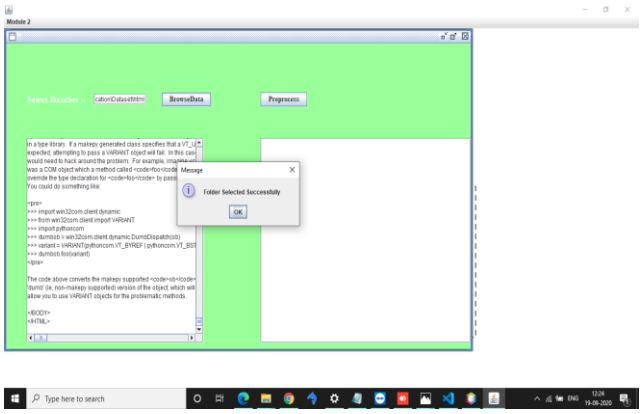

Figure 2: Data Selection

In figure 2 data is selected from web.

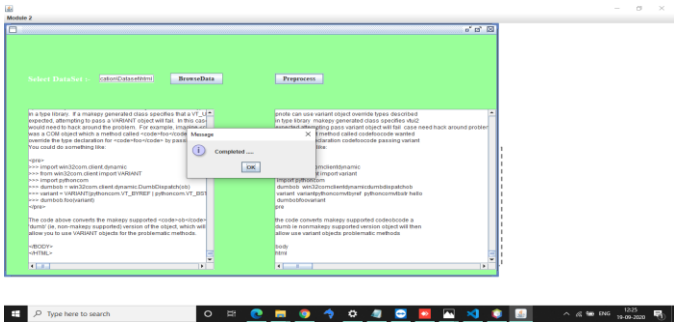

Figure 3: Data Preprocessing

In figure 3 data is preprocessed through different stages.

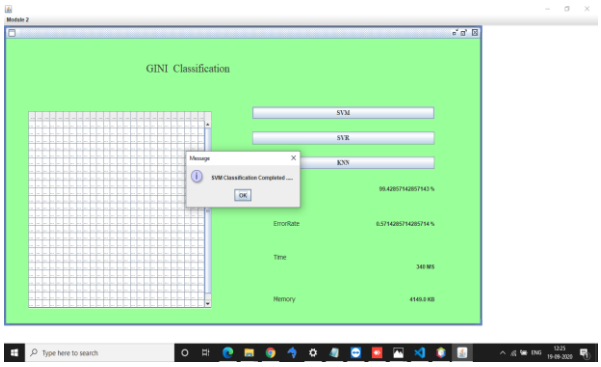

Figure 4: Gini Index Classification

The figure 4 shows the implementation Gini Index based feature selection and classification with different classifiers. The figure 5 shows the calculation of information gain based feature selection technique with different classifiers.

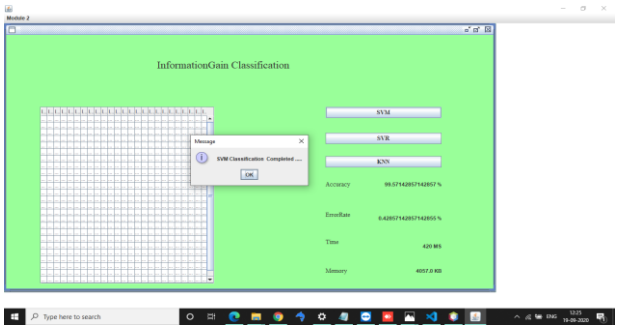

Figure 5: Information Gain Classification

V. RESULT ANALYSIS

In this experiment the selected feature selection technique is experimented with three classifiers i.e. SVM, SVR and k-NN. The key aim of this experiment is to identify the suitable classifier which is able to work with the defined feature selection techniques.

1) Accuracy- That can be estimated utilizing the proportion of all-out accurately characterized and the all-out patterns to be grouped.

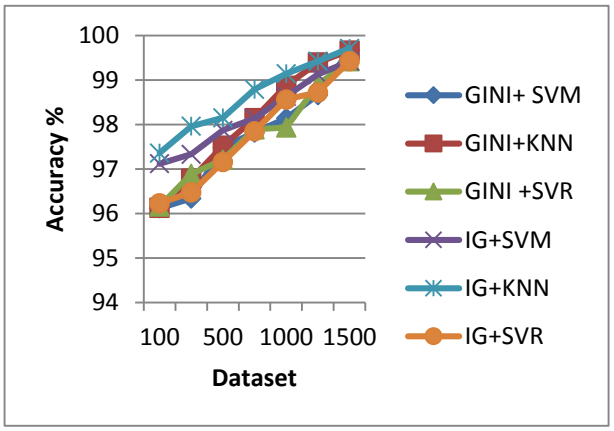

Figure 6: Accuracy

The accuracy of the techniques is notified in Y-axis in terms of percentage (%). Additionally the X-axis shows the size of dataset for experimentation.

2) Error rate- The error rate of a calculation demonstrates the misclassification rate of the calculation as a presentation boundary. That is opposite of accuracy and reverse in their processes.

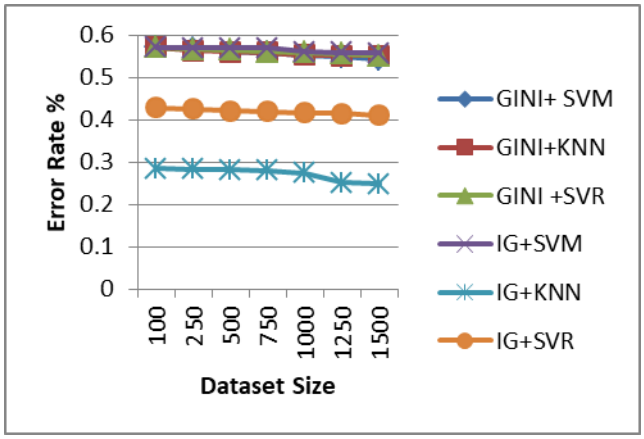

Figure 7: Error rate (%)

3) Time Consumption- The time consumption is measured here in terms of milliseconds (MS). Basically that is the total amount of time which taken to complete the training of the target algorithm.

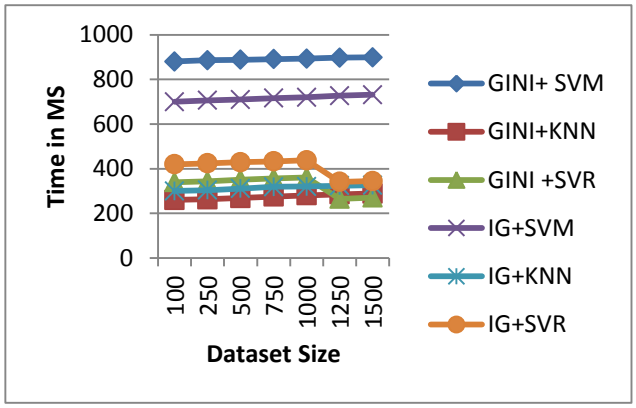

Figure 8: Time consumption (MS)

4) Memory Usage -The amount of total memory utilized for execution of an algorithm is measured as the memory consumption or usages

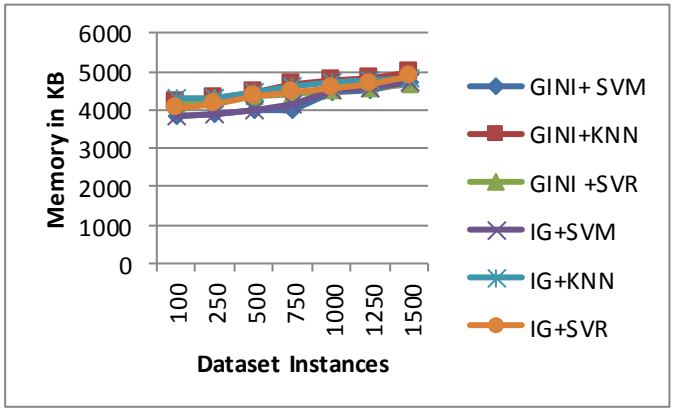

Figure 9: Memory Usage (KB)

Comparison of Feature Extraction Techniques with different classifiers

According to the data in the table 5.10 the accuracy is the highest of the Information Gain with the KNN classifier. Same is with the error rate and time consumption. The memory usage of GINI Index technique with SVM classifier is the least among all.

Table 5: comparison summary

Parameters	GINI+SVM	GINI+KNN	GINI+SVR	IG+SVM	IG+KNN	IG+SVR
Accuracy	97.74	98.07	97.77	98.23	98.64	97.77
Error Rate	0.557	0.558	0.561	0.565	0.271	0.419
Time	890	274	355	715	314	404
Memory	4199	4603	4426	4229	4568	4456

VI. CONCLUSION

Mining always produce fruitful results, however it is gold mining or data mining. In this generation data is expensive then gold. The processing of raw data and obtaining fruitful results are known as data mining. Similarly the data mining algorithms when applied on web data then it is called web data mining or web mining. The web mining can be classified into three type's web content mining, web usages mining and structure mining.

The web content mining techniques are employed on web pages to consume HTML (Hypertext Markup Language) or other web documents to identify the meaningful contents. Similarly, web usages mining explores the domain of hidden knowledge in web access log files. In this experiment we are use the machine learning techniques which can works in supervised learning manner for classifying the computed features. Therefore in this experimental analysis we include three popular and accurate machine learning algorithms (i.e. SVM, SVR, and KNN) for learn and classify the contents of web pages according to the predefined domains. The comparative study demonstrate the SVM and SVR classifier are providing higher accuracy as compared to k-NN additionally the k-NN algorithm found much costly for running time and memory usage.

REFERENCES

[1] D. Li, Y. Zhao, Y. Li, "Time-Series Representation and Clustering Approaches for Sharing Bike UsageMining", IEEE Access, Vol 7, Page. 177856-177863, 2019.
[2] A. Dutt, M. A. Ismail, T. Herawan, "A Systematic Review on Educational Data Mining", IEEE. Translations and content mining, Vol 5, Page. 15991-16005, 2017.
[3] A. Kumar and R. K. Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey", International Research Journal of Engineering and Technology, Vol 3, Issue: 12 Page. 1543-1547, 2016.
[4] F. Wu, X. Y. Jing, J. Zhou, Y. Ji, C. Lan, Q. Huang, R. Wang, "Semi-supervised Multi-view Individual and Sharable Feature Learning for Webpage Classification", In Proceedings of the 2019 World Wide Web Conference WWW '19, (San Francisco, CA, USA) pp.3349- 14 3355, 2019.
[5] S. Kumar, X. Gao, I. Welch, M. Mansoori, "A Machine Learning based Web Spam Filtering Approach", 30th International Conference on Advanced Information Networking and Applications, (IEEE), pp. 973-980, 2016.
[6] A. P. G. Plaza, V. Fresno, R. Mart´ınez, A. Zubiaga, "Using Fuzzy Logic to Leverage HTML Markup for Web Page Representation", IEEE Transactions On Fuzzy Systems, pp. 1-22, 2016.
[7] H. Khalifi, A. Elqadi, Y. Ghanou, "Support Vector Machines for a new Hybrid Information Retrieval System", The First International Conference on Intelligent Computing in Data Sciences, (Elsevier), pp. 139–145, 2018.
[8] M. O. Samuel, A. I. Tolulope, O. O. Oyejoke, "A Systematic Review of Current Trends in Web Content Mining", IOP Conf. Series: Journal of Physics, Vol.1299, pp. 1-14, 2019.
[9] H. Park, H. C. kwon, "improved GINI index algorithm to correct feature selection Bias in Text Classification", IEICE Trans. INF. & SYST, VOL E94D, No 4 April 2011.
[10] A. I. Pratiwi, Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis", Hindawi Applied Computational Intelligence and Soft Computing, Article ID 1407817, 5 pages, 2018.