Learning Approach for Automatic Web Usage Mining Using Multi-Label K-Nearest Neighbor (ML-KNN)

Er. Amandeep Kaur[1], Er. Sukhpreet Kaur[2]
[1]M.Tech Scholar, [2]Assistant Professor
Department of Information Technology[1], Department of Computer Science and Engineering[2]
SUSCET Tangori, Punjab, India

Abstract **-** this paper proposed a multiple-level strategy to automatically predict multiple labels of any web news on Real-Time basis. The TF-IDF technique has been adopted to represent user RSS address feed data in the Vector Space Model. An approach has been applied to the news dataset of different tags or categories searched by any user on-line. A comparison has been performed between KNN and ML-KNN techniques on the user RSS address feed dataset by creating multi-label classifiers.

Keywords **-** Data Mining, K-Nearest Neighbor, Web usage Data Mining, multi-label K-nearest Neighbor (ML-KNN), Weka.

I. INTRODUCTION

Data mining is the method of extraction of meaningful data from a larger set of any raw data, to discover hidden patterns in data; their relationships and summarize the data in a useful and understandable structure to the data users.
Data mining techniques another application is Web usage mining which is used to figure out meaningful data from any website. The Web Data Mining is an approach used to creep Adeniyi and Yongquan [3] adopted KNN algorithm in their paper to design and develop an automatic web usage data mining and recommendation scheme on present client nature during their click stream data on RSS reader websites.

The formation of Multi-label learning technique is from the search of text classification issue, in which every document can relate to various prefixed areas at the same time. The training group is formed of instances in multi-label learning; where everyone belonged to a group labels, and job is to foresee the label groups of hidden instances during the inspection of training instances with well-known label groups. A multi-label lazy learning known as ML-KNN is formed of the K-Nearest-Neighbour (KNN) method.

Representation of the paper in different sections as: Section 2 contains Literature review. Section 3, ML-KNN algorithm is briefly introduced. In the Section 4, Results of comparison between KNN and ML-KNN techniques are presented. In the Section 5, proposed work has been concluded. Finally Section 6 presents the Future scope.

through different web pages to gather required useful information. It was first introduced by Etzioni in his paper "The World Wide Web: Quagmire or Gold Mine" [1] in 1996 after that researchers world has been moved towards this important research area.

The RSS (Really Simple Syndication) websites were developed to access daily news on-line across the world. The first version of RSS known as RDF Site Summary was beloved to create at Netscape by Dan Libby and Ramanathan V. Guha to use on the My.Netscape.Com portal [2] it was released in March 1999. The goal of the study is to design and develop an on-line, automatic and Real-Time web usage data mining system based on TF-IDF technology. The system is able to predict categories of news automatically.

The K-Nearest-Neighbor (KNN) approach is basically for on-line and real-time identification of user or guests click current information, coordinating it to a distinct group of users. Mostly, for Pattern identification; K-NN approach is used. It is a type of lazy learning method. The *k*-NN method is the simplest method among all the machine learning methods.

II. LITERATURE REVIEW

Adeniyi and Yongquan et al. [3] presented a study to design and develop an automatic web usage data mining and recommendation scheme on present client nature during their click stream data on RSS reader websites, as to give them suitable information without explicitly asking for it.
Arya et al. [4] presented an automatic recognition method that has been recommended for a classification of news web pages, uses classification rules based on a combination of content, structure and uniform resource locator (URL) attributes. In this work, they gathered news web documents from 10 different news websites and used Naïve Bayes algorithm to differentiate news articles from non-news articles.
Kwon et al. [5] proposes an algorithm of RSS-based indoor localization. For wireless sensor networks that is combined with PDR location tracking. The main aim of the study is by using PDR technique to compensate the NLOS localization problem, while checking the assembled problem of PDR by

using RSS-based localization technique in LOS circumstances.

Remya et al. [6] presented a survey on web mining strategies used for mining different forms of data on existed Web. For data extraction from web pages, it presented a technique called top-k web pages that describes top-k instances of any specific concerned topic which is very handy in search or fact answering systems.

Mehta et al. [7] proposed a work of discovery of user pattern in accessing website using web log history. The goal of this research was to discover user access patterns based on the user's session and behavior. In their work they applied a collaboration of clustering and association rule mining for pattern findings.

Dharmarajan, Dorairangaswamy et al. [8] introduced a method of web usage mining to figure out the nature of website clients during the development of data mining of web approach information. They used FP-growth approach to obtain periodic access patterns from the web log data and provide beneficial information about the user's concern.

Zadrozny et al. [9] concluded their work with actual ideas for future research in multi-label feature selection which is derivative of categorization and analysis.

Zhang, Zhou et al. [10] presented a multi-label lazy learning technique titled ML-KNN in his paper, ML-KNN is a derived version of classical K-nearest neighbor (KNN) technique. Specifically, for every hidden instance, basically its K nearest neighbors in the training group is analyzed. They presented three experiments of various physical world multi-label learning issues that are Yeast gene functional analysis, natural scene classification and automatic web page classification, which concluded the ML-KNN, attains remarkable efficiency to few absolute multi-label learning approaches.

Younes, Abdallah, Denoeux et al. [11] describes a new technique for multi-label classification issues derivative of Bayesian form of the K-nearest neighbor (KNN), and taking into worth the dependent relations in labels. Tests on standard datasets present the value and the performance of the presented approach related to some another approaches.

Dario Antonelli et al. [12] proposed a data analysis scheme to recognize observation tracks followed by patients. A multiple-level collection technique is used to recognize a group of data with a variable allocation. They adopted TF-IDF scheme to represent patient examination data.

Federico and Pier et al. [13] presented a study of the recent developments in web mining area that is taking high attention from the Data Mining community.

III. ML-KNN ALGORITHM

For comfort, various documentations are received before showing ML-KNN. Here *x* is an instance and its related label set $Y \subseteq$ Y; assume KNNs are taken in the ML-KNN approach. Assume category vector of *x* be $\overrightarrow{y_x}$, where its *l*th component $\overrightarrow{y_x}(l)$ ($l \in$ Y) holds the value of 1 if $l \in$ Y and 0 alternatively. Further, assume *N(x)* stands for the group of KNNs of *x* analyzed in the training set. Thus, on the basis of label sets of these neighbors, a *membership counting* vector can be shown as

$$\overrightarrow{C_x}(l) = \sum_{a \in N(x)} \overrightarrow{y_a}(l), \qquad l \in \mathrm{Y}, \tag{1}$$

where $\overrightarrow{C_x}(l)$ counts the number of *x*'s neighbors relating to the *l*th class.

Consider every test instance *t*; ML-KNN initially analyzes its KNNs *N (t)* in the training group. Assume H_1^l be the event that *t* has label *l*, where as H_0^l be the event that *t* has not label *l*. In addition, assume E_j^l ($j \in \{0, 1, \ldots, K\}$) show the event that, with the KNNs of *t*, there are specially *j* instances that have label *l*. So, on the basis of membership counting vector $\overrightarrow{C_t}$, the category vector $\overrightarrow{y_t}$ is driven using the below MAP principle:

$$\overrightarrow{y_t}(l) = \arg\max_{b \in \{0,1\}} P\left(H_b^l \middle| E^l_{\overrightarrow{C_x}(l)}\right), l \in Y. \tag{2}$$

By using the Bayesian rule, Eq. (2) can be revising as

$$\overrightarrow{y_t}(l) = \arg\max_{b \in \{0,1\}} \frac{P(H_b^l) P\left(E^l_{\overrightarrow{C_t}(l)} \middle| H_b^l\right)}{P\left(E^l_{\overrightarrow{C_t}(l)}\right)}$$

$$= \arg\max_{b \in \{0,1\}} P(H_b^l) P\left(E^l_{\overrightarrow{C_t}(l)} \middle| H_b^l\right). \tag{3}$$

*Algorithm is taken from paper "ML-KNN: A lazy learning approach to multi-label learning" by Zhang, Zhou [10]

$[\vec{y}_t, \vec{r}_t]$=ML-KNN(T, K, t, s)

%Computing the prior probabilities $P(H_b^l)$

(1) **for** $l \in \mathcal{Y}$ **do**

(2) $P(H_1^l) = (s + \sum_{i=1}^{m} \vec{y}_{x_i}(l)) / (s \times 2 + m)$; $P(H_0^l) = 1 - P(H_1^l)$;

%Computing the posterior probabilities $P(E_j^l | H_b^l)$

(3) **Identify** $N(x_i)$, $i \in \{1, 2, \ldots, m\}$;

(4) **for** $l \in \mathcal{Y}$ **do**

(5) **for** $j \in \{0, 1, \ldots,$ K$\}$ **do**

(6) $c[j] = 0$; $c'[j] = 0$;

(7) **for** $i \in \{1, 2, \ldots, m\}$ **do**

(8) $\delta = \vec{C}_{x_i}(l) = \sum_{a \in N(x_i)} \vec{y}_a(l)$;

(9) **if** $(\vec{y}_{x_i}(l) == 1)$ **then** $c[\delta] = c[\delta] + 1$;

(10) **else** $c'[\delta] = c'[\delta] + 1$;

(11) **for** $j \in \{0, 1, \ldots,$ K$\}$ **do**

(12) $P(E_j^l | H_1^l) = (s + c[j]) / (s \times (K + 1) + \sum_{p=0}^{\mathrm{k}} c[p])$;

(12) $P(E_j^l|H_1^l) = (s + c[j])/(s \times (\kappa + 1) + \sum_{p=0}^{k} c[p])$;

(13) $P(E_j^l|H_0^l) = (s + c'[j])/(s \times (\kappa + 1) + \sum_{p=0}^{k} c'[p])$;

%Computing $\vec{y}_t$ and $\vec{r}_t$

(14) **Identify** $N(t)$;

(15) **for** $l \in \mathcal{Y}$ **do**

(16) $\vec{C}_t(l) = \sum_{a \in N(t)} \vec{y}_a(l)$;

(17) $\vec{y}_t(l) = \arg\max_{b \in \{0,1\}} P(H_b^l) P(E^l_{\vec{C}_t(l)}|H_b^l)$;

(18) $\vec{r}_t(l) = P(H_1^l|E^l_{\vec{C}_t(l)}) = (P(H_1^l) P(E^l_{\vec{C}_t(l)}|H_1^l))/P(E^l_{\vec{C}_t(l)})$
$= (P(H_1^l) P(E^l_{\vec{C}_t(l)}|H_1^l))/(\sum_{b \in \{0,1\}} P(H_b^l) P(E^l_{\vec{C}_t(l)}|H_b^l))$;

Fig: 1 ML-KNN Pseudo Code

As in Eq. (3), in order to decide the category vector $\overrightarrow{y_t}$, all the data required is the prior probabilities $P(H_b^l)$ ($l \in$ Y, $b \in \{0, 1\}$) and the posterior probabilities $P(E_j^l|H_b^l)$ ($j \in \{0, 1, \ldots, K\}$). In reality, these prior and posterior probabilities may all be straightly supposed from the training set on the basis of periodic counting.

Figure 1 presents the whole information of ML-KNN. *T* is the training set *T* = {*(x*1, *Y*1*), (x*2, *Y*2*), . . . , (xm, Ym)*} *(xi* ∈ X, *Yi* ∈ Y*)* and the input variables K, *t* and the output variable $\overrightarrow{y_t}$ meanings are the same as explained earlier. Moreover, the input argument *s* is a smoothing parameter controlling the strength of consistent prior. $\overrightarrow{r_t}$ Is an actual-valued vector calculated to rank labels in Y, where $\overrightarrow{r_t}(l)$ corresponds to the posterior probability$P\left(H_1^l \middle| E^l_{\overrightarrow{C_t}(l)}\right)$. Figure 1 show that, on the basis of multi-label training instances, steps (1) and (2) are estimated the prior probabilities $P(H_b^l)$. Whereas from Steps (3) to (13) estimates the posterior probabilities $P(E_j^l|H_b^l)$, where $C[j]$ is being used in every iteration of *l* which counts the number of training instances with label *l* whose KNNs have same *j* instances with label *l*. Therefore, $C'[j]$ counts the number of training instances without label *l* whose KNNs have exactly *j* instances with label *l*. At the end, by using the Bayesian rule, steps from (14) to (18) calculates the method's outputs on the basis of supposed probabilities.

IV. IMPLEMENTATION AND RESULTS

The comparison of KNN and ML-KNN is performed on the basis of three parameters. Those are:

A. Average Precision.
B. Error Rate.
C. Execution Time.

To perform this work, dataset has been adopted from the work of Adeniyi, Yongquan [3]. [3] Performed experiment on the basis of present user's behavior through their clicks stream data on RSS reader websites. But in this proposed work, experiment is performed on the basis of multi-labels (tags or categories) to which news belongs to. To perform final results, code is written in MATLAB and the final dataset is created with the help of WEKA tool. For the final dataset creation some steps to be performed. Steps are as follows:

Sentence Splitting - The first step involved is sentence splitting i.e. the splitting of string into words. Identifying sentence boundaries in a document is not a smaller task.

Tokenization - Tokenization of words meant to label the individual words or word parts. When the file is taken as input, after splitting it is divided into tokens which are labeled individually. Tokenization is an important task because many succeeding components need tokens clearly identified for analysis.

Tokenization of words meant to label the individual words or word parts. When the file is taken as input, after splitting it is divided into tokens which are labeled individually. Tokenization is an important task because many succeeding components need tokens clearly identified for analysis.

Stop Word Filtering - There are a lot of words that do not have any meaning and that can be removed from the input file. Words like "the", "and", "or", "if", such type of words do not signify anything so they can be removed. Removal of such words refers to stop word filtering and it also improves performance of the system. The basic approach for doing this task is to remove all words that occur on a list of common words. One more approach is to remove all the words that appear with high frequency across most of the documents. Stop words are not useful in any way so they create noise in data and must be removed. The stop word filter will remove token annotations from the documents.

Stemming - A stemming algorithm is a process of linguistic normalization. In this process the variations of words are reduced to a common form.

TF-IDF - Using the "Term Frequency Inverse Document Frequency" i.e. TF-IDF as a feature selection method. TF-IDF is basically a mathematics technique of numerical statistic. TF-IDF is defined as the multiplication of two numerical statistics i.e. TF and IDF. It detect the recurrence of words in a document by finding the value of appropriate words by an inverse ratio of the recurrence of word in a document verses the total percentage of documents the words arises in them. TF-IDF produces high percentage if the words are occurs various times in a single document.

Table 1 Calculated Values of KNN and ML-KNN classifier

Classifiers	Average Precision	Error Rate	Execution Time
KNN	0.6600	21.5123	0.8784
ML-KNN	0.8041	19.5933	0.6687

Table 1 shows the final calculated results of KNN and ML-KNN classifiers for the avrage precision, error rate and execution time and results shows that ML-KNN classifier performance is excelling than KNN in terms of all the three parameters being used in the work. Below is the graphical representation of the results. In the Figure 2 the avrage precision is shown between KNN and ML-KNN classifier. In the Figure 3 the error rate of KNN and ML-KNN represented and in the Figure 4 the execution time is ploted on the graphs. These graphs are created in MATLAB.

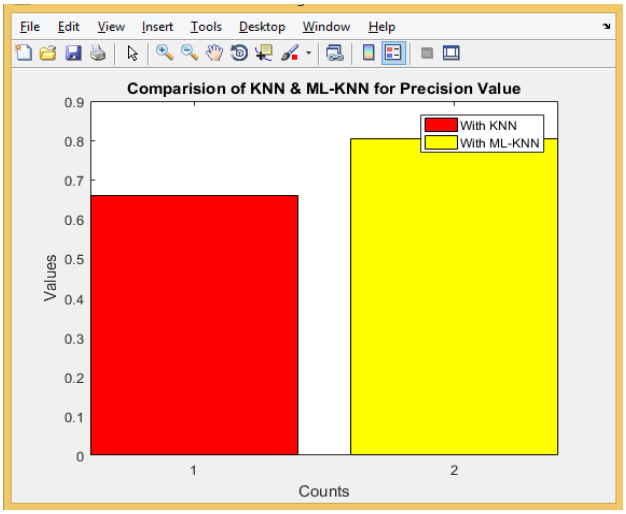

Fig: 2 Precision values of KNN and ML-KNN classifiers

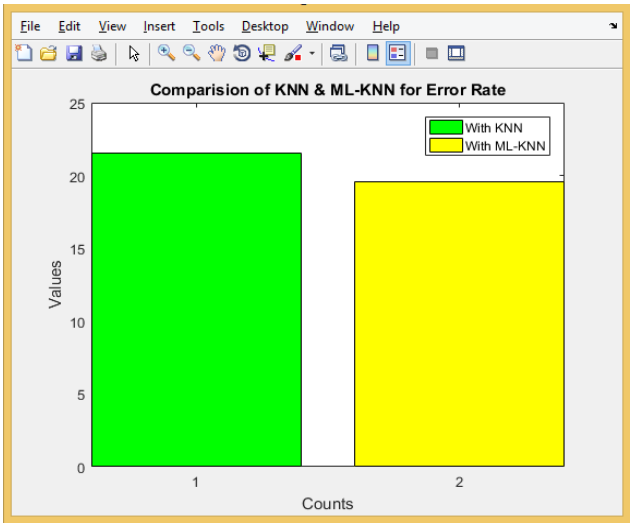

Fig: 3 Error Rate of KNN and ML-KNN classifiers

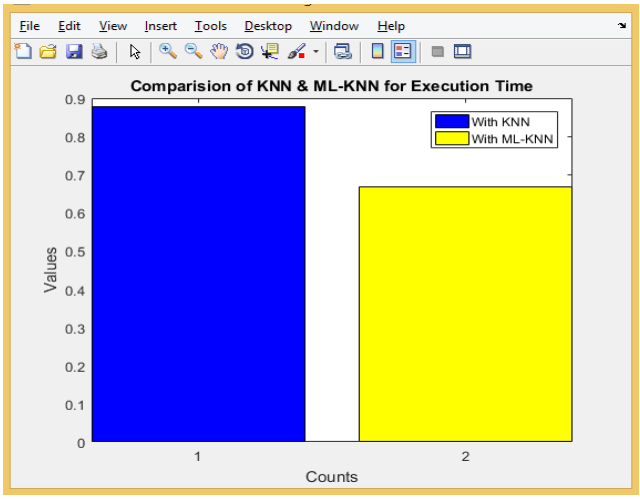

Fig: 4 Execution Time of KNN and ML-KNN classifiers

V. CONCLUSION

The proposed work is on the basis of automatic Real-Time web usage data mining system. The system performs multi-label classifiers on news tags of user RSS address feeds. The system is able to automatically predict multi-labels of any web news on Real-Time basis.

The proposed work result shows that an automatic web usage mining for multi-labels ML-KNN classifier performance is excelling than KNN in terms of all the three parameters being used in the work i.e. Precision value, Error rate and Execution time. ML-KNN works better when there is a need of multi-label classifiers.

VI. FUTURE SCOPE

In future the proposed scheme can be tested for other datasets. More research also can be carried out on many other data mining techniques (like Neural Network, SVM), comparing the result with this mode.

REFERENCES

[1]. Oren Etzioni, "The World Wide Web: quagmire or gold mine?" Department of Computer Science and Engineering University of Washington, Seattle, WA, Nov.'96.

[2]. Hines, Matt (1999-03-15). "Netscape Broadens Portal Content Strategy" *Newsbytes*.

[3]. D.A. Adeniyi, Z. Wei, Y. Yongquan "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method" 2210-8327 @ 2015.

[4]. Sanjay K. Dwivedi, Chandrakala Arya "News Web Page Classification Using Url Content and Structure Attributes" October 2016.

[5]. Hyunmin Cho, Younggoo Kwon "RSS-based indoor localization with PDR location tracking for wireless sensor networks" Volume 70, Issue 3, March 2016.

[6]. Asha Joy, Remya R "Techniques for Web Mining of Various Forms of Existence of Data on Web: A Review" Volume 3, Issue 1, January 2015.

[7]. Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery" Volume 2, Issue 1, June 2013.
[8]. K. Dharmarajan, Dr. M. A. Dorairangaswamy "Web Usage Mining: Improve The User Navigation Pattern Using Fp-Growth Algorithm" Volume -3, Issue-4, August 2016.
[9]. Pereira, R.B., Plastino, A., Zadrozny, B. et al. "Categorizing feature selection methods for multi-label classification "Artif Intell Rev (2016).
[10]. Min-Ling Zhang, Zhi-Hua Zhou "ML-KNN: A lazy learning approach to multi-label learning" 15 December 2007.
[11]. Zoulficar Younes, Fahed Abdallah, And Thierry Denoeux "Multi-label classification algorithm derived from *k*-nearest neighbor rule with label dependencies".
[12]. Dario Antonelli, Elena Baralis, Giulia Bruno, Tania Cerquitelli, Silvia Chiusano, Naeem Mahoto "Analysis of diabetic patients through their examination history" Volume 40, Issue 11, 1 September 2013, Pages 4672-4678.
[13]. M.F. Federico, L.L. Pier, Mining interesting knowledge from weblog: a survey, J. Data Knowledge Eng. 53 (2005) (2005) 225–241.

I am Amandeep Kaur. I received B-Tech degree in Information Technology from Sachdeva Engineering College for Girls, Gharuan under Punjab Technical University, Jalandhar, Punjab, India, in 2008. Now, I am pursuing M-Tech in IT (Part Time) from Shaheed Udham Singh College of Engineering & Technology, Tangori (Mohali) under I K Gujral Punjab Technical University, Jalandhar, Punjab, India. I have 3.6 years of experience as a web developer in IT industry. Currently, I am done with my M-Tech thesis research. My research interest includes web data mining, automated web usage mining using multi-label K-nearest Neighbor (ML-KNN) on RSS user address feeds by using TF-IDF technique.