Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/3793
DOI URL: http://dx.doi.org/10.21474/IJAR01/3793
ISSN NO. 2320-5407

RESEARCH ARTICLE

CLASSIFICATION OF WEB DOCUMENTS USING HYBRID FEATURE SELECTION.

V. David Martin[1] and Dr. T. N. Ravi[2]
1. Research Scholar, Periyar E.V.R College (Autonomous), Trichy.
2. Assistant Professor, Periyar E.V.R. College (Autonomous), Trichy

Manuscript Info

Manuscript History

Received: 01 February 2017
Final Accepted: 10 March 2017
Published: April 2017

Key words:-
Particle Swarm Optimization,
Relative Reduct

Abstract

Knowledge discovery and data mining is a process of retrieving the meaningful knowledge from the raw data, using different techniques. Therefore, text mining is a sub domain of knowledge discovery from the text data. Web mining is a one class of data mining. Web Mining is a variation of data mining that distills untapped source of abundantly available free textual information. The need and importance of web mining is growing along with the massive volumes of data generated in web day-to-day life. Feature selection is an effective technique for dimension reduction and an essential step in successful data mining applications. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. In this paper, a hybrid feature selection is proposed. The Relative Reduct and Particle Swarm Optimization Technique are hybridized to reduce the size of the feature space

Introduction:-

In the modern days of technology text mining studies are advancing into next level due to mounting number of the electronic documents from a mixture of resources. The resources of unstructured and semi structured data includes the World Wide Web, Governmental Electronic Repositories, News Editorials, Genetic Directory , Depositories of Blog, Online Forums, Digital Libraries, Electronic Mail and Chat Rooms. Consequently, the appropriate categorization and knowledge detection from these sources and it marks a major role in the field for investigation.

Natural Language Processing (NLP), Information Mining, and Machine Learning methods work reciprocally in categorizing the determine patterns instinctively from the electronic documents. The primary objective of the text mining is to facilitate clients to extort information from textual resources and compacts with the maneuvers like, repossession, categorization (supervised, unsupervised and semi supervised) and recapitulation. In contrast, how these documents can be aptly interpreted, presented and classified. In view of that, it consists of numerous challenges, like proper explanation to the documents, with appropriate file demonstration, dimensionality diminution to grip algorithmic concerns [1]. Moreover a suitable classifier jobs are occupied to accomplish good overview and evade over-fitting. Mining, incorporation and categorization of electronic data from miscellaneous sources and knowledge discovery have been composed from these documents to channel it for the research societies.

At present, the web is the chief source for the text documents, the quantity of textual data existing to us is constantly mounting, and approximately 80% of the data of an organization is piled up in unstructured textual format [2] like

Corresponding Author:- V. David Martin.
Address:- Research Scholar, Periyar E.V.R College (Autonomous), Trichy.

reports, email, views and news etc. The [3] exhibits that just about 90% of the Global data is apprehended in unstructured formats, as a result Information intensive business processes stipulate that we surpass since simple document retrieval to knowledge discovery. The requirement of automatically retrieval of useful knowledge with the colossal amount of textual data is used to facilitate the support of human investigation is fully perceptible [4].

The movement of the market is depended on the information of the online news articles, reactions, and proceedings is turned out to be an budding theme for investigation in the field data mining and text mining [5].To establish the outcome of modern methods for text classifications are elucidated in [6] which three problems were highlighted: documents demonstration, classifier erection and classifier assessment. Therefore, generating an information structure that can signify the data, and creating a classifier that can be utilized to visualize the class label of a document with high accuracy, develops into the major issues in text categorization.

Related Works:-

In the paper, the authors Naw, Naw, and Ei Ei Hlaing [7] used the Episode Rules for the application of finding keywords and key phrases, discovering grammatical rules and collections by using bag of words and word positions. The authors Adeva, JJ García, et a [8] utilizes TFIDF and Naïve Bayes classification algorithm for the text classification using Bag of Words. Cohen [9] in the paper, used Propositional Rule based system and Inductive Logic Programming for the text classification by Relational. The authors Menaka. S and N. Radha [10] in the paper utilize the TFIDF, Decision Trees, Naïve Bayes, Bayes Net and Support Vector Machine for the text categorization by using Bag of Words phrases. Koplenig, Alexander, et al [11] finding patterns between concept distribution in textual data by relative entropy using concept categories. qbal, Farkhund, et al. [12] finding patterns across terms in textual data by using Association Rules. Taghandiki, Kazem, Ahmad Zaeri, and Amirreza Shirani [13] used Naïve Bayes classification algorithm for extracting key phrases from text documents using Phrases and their positions. Tobon-Mejia, Diego Alejandro, et al [14] used Hidden Markov Modes for Learning Extraction Models using Bag of Words. The authors Grimmer, Justin, and Brandon M. Stewart [15] utilized Unsupervised statistical method for hierarchical clustering. Anami, Basavaraj S., Ramesh S. Wadawadagi, and Veerappa B. Pagi. [16] used Self Organizing Maps for the text and document clustering using Bag of Words with n-grams.

Relative Reduct Feature Selection Technique:-

In subsequent to characteristic mining, the significant procedure in pre-processing of manuscript classification is attribute selection. It is employed to create vector space, which progress the scalability, competence and exactness of a text classifier. In most cases, a good feature selection method should be considered the domain and algorithm features [17]. The central design of FS is to opt for compartment of features from the novel documents. Moreover, FS is acted upon by keeping the words with highest score in according to prearranged measure of the significance of the word. The selected feature conserves original substantial meaning and affords a better understanding for the information and learning process [18]. For text classification, the key issues are the elevated facets of the feature space. Nearly each text domain has large quantity of features, the majority of these features are not significant and complimentary for text categorization task, and still the various noise features may stridently diminish the classification accuracy [19]. Therefore, FS is generally deployed in text categorization to shrink the dimensionality of feature space and progress the competence and exactness of classifiers.

There are generally two kinds of feature selection techniques in machine learning; wrappers and filters. Wrappers utilized in the classification precision of some learning algorithms as their assessment function. Since wrappers have to instruct a classifier for every feature subset to be assessed, they are generally much more time consuming. In particularly, when the number of features is high. Accordingly, wrappers are by and large not suitable for text categorization. At the same time as opposed to wrappers, filters act upon FS autonomously of the learning algorithm that determine to make use of the selected features. With the intention to assess a feature, filters exploit an assessment metric that calculates the capacity of the feature to distinguish each class [20]. In text categorization, a text manuscript may moderately match many categories. Therefore, we necessitate discovering the greatest identical group for the text document. The expression (word) regularity/converse document frequency (TF-IDF) technique is generally used to weight every word in the manuscript document according to how characteristic it is. In further expressions, the TF-IDF technique precincts the pertinent amongst words, text data and fussy classifications.

In Relative Reduct algorithm we find out the degree of relative dependency after removing the attributes from the set. If an attribute is removed and it causes the value of relative dependency to be one then that attribute is eliminated otherwise it is put in the core reduct. The process is repeated again and again till the value becomes one. The algorithm is explained below:

Relative Reduct Algorithm:-
Input: Original Dataset,
D the set of all conditional features; Y- Conditional attribute from R.
Step 1: D ← {list of all conditional features}
Step 2: Now select the conditional attribute Y from D.
Step 3: Calculate the relative dependency of the feature.
Step 4: If relative dependency of the attribute is one then eliminate the attribute, Go to step 2.
Step 5: Else Add that feature in Reduct set.
Step 6: Stop

Output:- Reduct dataset
The relative reduct algorithm endeavors to ascertain a reduct without completely creating every single possible subset. It begins off with a vacant set and includes turn, each one in turn, those features that outcome in the best increment in the rough set dependency metric, until this delivers its most extreme conceivable quality for the dataset.

Particle Swarm Optimization:-
Particle Swarm Optimization (PSO) [21] is based on the social behavior associated with bird's flocking for optimization problem. A social behavior pattern of organisms that live and interact within large groups is the inspiration for PSO. The PSO is easier to lay into operation than Genetic Algorithm. It is for the motivation that PSO doesn't have mutation or crossover operators and movement of particles is effected by using velocity function [22]. In PSO, every particle alters its own flying memory and its partner's flying inclusion keeping in mind the end goal to flying in the search space with velocity.

The best-fit particle of the entire swarm [23] influences the position of each particle. Each individual particle j $\in [1 \dots m]$ where m > 1, has current position in search space s_j, a current velocity u_j and a personal best position $p_{b,j}$ where j is the smallest value determined by objective function o. By using $p_{b,j}$ the global best position G_b is calculated, which is the buck value obtained by comparing all the $p_{b,j}$

The $p_{b,j}$ is calculated by using the formula

$$p_{b,j}=\begin{cases} p_{b,j}\ if\ (y_j) > p_{b,j} \\ y_j\ if\ f(y_j) \le p_{b,j} \end{cases}$$

The formula used to calculate Global Best Position G_{best} is

$$G_b = \{\min\{p_{b,j}\}, where\ j \in [1, \dots, m] where\ m > 1$$

Velocity can be updated by using the formula

$$u_j^{j+1} = wu_j(t) + s_1 i_1[y_j(t) - y_j(t)] + d_2 i_2[g(t) - y_j(t)]$$

where $u_i(t)$ is the velocity and w, s_1 and s_2 are used supplied co-efficient. The i_1 and i_2 are random values $y_j(t)$ is the individual best solution, g(t) is the swarm's global best candidate solution. $wu_j(t)$ is known as inertia component. Inertia component value lies between 0.8 and 1.2. Lower the values of inertia component, it speeds up the convergence of swarm to optima. But higher value encourages the exploration of entire search space. $s_1 i_1[y_j(t) - y_j(t)]$ is known as cognitive component.

Proposed Hybrid Feature Selection Algorithm:-
The existing Particle Swarm Optimization (PSO) and Relative Reduct (RR) algorithm are combined to develop the Hybrid Feature Selection Algorithm. The process of producing the optimal data set by the hybrid Feature selection algorithm is as follows: At the preliminary level, the initialization of Inertia (ω), Decision Features (Df) and Conditional feature (Cf), Social and Cognitive components (s1 & s2), Swarm Size (SS) and random values (i1 & i2) to be done. The Cognitive and Social (s1 & s2) values are initialized as 2.0. This is because if the value of s1 and

s2 is low, it will roam far from the target region. If the value is high, there will be an abrupt movement of particles towards the target. The random values of i1 and i2 lie between 0 and 1. The random values are set to be 0.2. Usually, the inertia value too lies between 0 and 1. If the inertia value is high, the particles will have high exploration capacity. If the inertia value is low, the particles will have high exploitation capacity. The inertia is initialized to 0.33 to have strong exploitation capability. The swarm size is fixed to be 20.

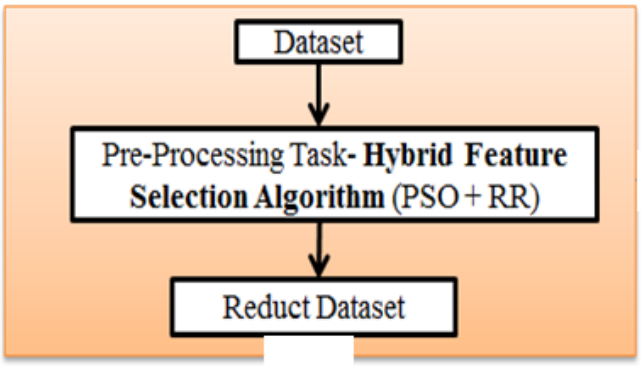

Figure 1:- Proposed Hybrid Feature Selection Framework.

After initializing the swarm size S, the data set is set to be a null or empty data set R ← { }. The empty data set R is then stored in T. After storing the empty data set in T, the conditional features are checked with empty set and initialized that $\forall_x \in (Cn - R)$. Then the Data Fitness Value (DFV) is calculated using the Griewangk function. This is because the Griewangk function has widespread local minima. The DFV is compared with best of the particle (p_b). If DFV is better than $P_{b,}$ then the P_b is set with current value of DFV. Then the particle with best fitness value is chosen by comparing $P_{b,}$ which is referred as the Global Best (G_b). Then the classification accuracy of the particle is compared with the Decision Feature. If the accuracy of the feature has value greater than the Decision Feature, then the feature is stored in T and passed to empty set R. Now, the R has a feature with higher classification accuracy. Then the velocity is calculated and updated by using the equation

$$U_{ud} = U_{ud} + S1i1(p_{jd} - y_{jd}) + S2i2(p_{gd} - y_{jd})$$
$$Y_{jd} = Y_{jd} + U_{jd}$$

$$V_{ud} = V_{ud} + c1r1(p_{id} - x_{id}) + c2r2(p_{gd} - x_{id})$$
$$X_{id} = X_{id} + V_{id} \quad (1)$$

This process is repeated until the classification accuracy of the reduct set R based on decision feature D is equal to classification accuracy of conditional feature based on D. The final output is the optimal reduct data set. The pseudo code for the hybrid Feature Selection Algorithm is as follows:

Pseudo code: Hybrid Feature Selection Algorithm:-
Input:- Data set
Algorithm:-

```
Initialize Cn, Df,SS, s1,s2, i1,i2, ω
For each particle i in S do
        {
                R ← { }
        }
        do
                T ← R
                ∀x ε (Cn – R)
                Calculate DFV
                If DFV > Pb
                        Set Pb = DFV
                If Pb > Gb
                        Gb = Pb
```

If $\gamma_{RU}(y) > \gamma_T(D)$
$T \leftarrow R_U(Y)$
$R \leftarrow T$

Calculate and update velocity by using the equation 1
Until
If $\gamma_R(Df) == \gamma_C(Df)$
Return R
Output: Optimal Reduct Data set.

Information of the Dataset:-
The experimental setup is explained in this section. It wraps up facts in relation to the datasets so as to deploy and dissimilar preprocessing methods that were practiced. The software device and enclosed that are utilized, Hardware and software particulars of the mechanism, lying on the study was performed in.

The following table 1 represents the various text classification datasets. The table shows that the text classification datasets are available in the repository [24]. The original datasets are divided into a collection of N binary ones (one for each class label of the original dataset). Each one of these binary datasets consists of 100 attributes. Each data file in the dataset has the following structure:
@relation - Name of the Dataset
@attribute - Description of an attribute (one for each attribute)
@inputs - List with the names of the input attributes
@output - List with the names of the output attributes
@data - Starting tag of the data

Table 1:- Various Documents Datasets Characteristics.

Name	Number of Attributes	Examples	Labels
blogsGender-100	100	3232	2
OH0-100	100	1003	10
OH10-100	100	1050	10
OH15-100	100	913	10
OH5-100	100	918	10
Ohscale-100	100	11162	10
Ohsumed-100	100	13929	10
r10-100	100	12897	10

From the table 1, in this paper, last dataset is considered for the research methodology i.e.r10-100 dataset contains 10 files each have 100 attributes. The table 2 gives the description about the r10-100 dataset.

Table 2:- Files and number of attributes in r10-100 Document dataset.t

Name	Number of Attributes
acq	100
Corn	100
Crude	100
Earn	100
Grain	100
Interest	100
Money-fx	100
Ship	100
Trade	100
Wheat	100

The software and hardware used are as follows: Processor: Intel Core i3 CPU M350 @ 2.27 GHz RAM: 3.00 GB Operating classification: Windows 7 Ultimate. MATLAB R 2016 a.

Evaluation Criteria:-

Based on the algorithms of different feature selection, the performance of the categorization of document is evaluated by adopting the evaluation criteria. Here to validate the results of the proposed Hybrid Feature Selection Algorithm, the following parameters like Classification Accuracy, Kappa Statistic, Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Absolute Error (RAE), Root Relative Absolute Error (RRAE), True Positive Rate (TPR), False Positive Rate (FPR), Recall, Precision, Confusion Matrix and ROC (Receiver Operating Characteristic Curve) are considered. The average square difference between the outputs and targets is Mean Squared Error. If zero it means no error whereas lower values are better. The correlation between targets and output is measured by value called Regression R. The random relationship is indicated by 0 whereas close relationship is given by 1. The TPR against FPR is considered to plot the ROC curve. The obtained result is considered as good only when the ROC value areas are nearer to the value of 0.80 to 0.90. The predictive model performance is estimated by using Cross Validation technique. In the 10 fold cross validation, the data sets are divided into 10 sets in which 9 data sets are used for training and 1 is used for testing.

Experimental Result and Analysis:-

In the first step, the proposed Hybrid Feature Selection algorithm is used to reduce the size of the feature space or number of attributes. Table 3 represents the datasets total number of attributes and the results obtained by proposed Hybrid Feature Selection and existing methods like PSO and Relative Reduct.

Table 3:- Feature Reduction by Relative Reduct, Particle Swarm Optimization and Hybrid Method for various datasets.

Datasets	Number of Attributes	PSO	Relative Reduct	Hybrid Feature Selection (RR+PSO)
Acq	100	45	35	30
Corn	100	48	32	28
Crude	100	42	30	25
Earn	100	41	36	29
Grain	100	50	25	25
Interest	100	49	33	29
Money-fx	100	52	36	27
Ship	100	33	30	21
Trade	100	39	39	30
Wheat	100	40	28	22

Organizing the data into categories is the primary work of classification. In this work, the web document is classified into two categories as significant and no significant. The reduct set obtained by applying proposed Hybrid Feature Selection algorithm is subjected to the classification. The reduced set obtained from the given dataset and the quality of reduct set is determined based on the achieved classification accuracy.

Table 4:- Classification Accuracy of the Original Dataset, Relative Reduct, PSO, Hybrid Feature Selection Algorithm

Evaluation Metrics	Original data set	PSO	RR	Hybrid Feature Selection (RR+PSO)
Correctly classified instance	72.19	79.74	78.54	82.61
Kappa statistic	0.37	0.45	0.58	0.66
MAE	0.33	0.26	0.23	0.22
RMSE	0.47	0.44	0.38	0.35
RAE	67.47	52.57	47.09	46.25
RRAE	89.88	87.67	78.57	77.64
TPR	0.69	0.69	0.79	0.81
FPR	0.31	0.32	0.21	0.21
Precision	0.69	0.69	0.79	0.81
Recall	0.69	0.69	0.79	0.81
ROC Area	0.39	0.73	0.85	0.84

It is observed from the Table 4 that the increased accuracy level is achieved on the optimal data set obtained by applying Hybrid Feature Selection algorithm. Accuracy level obtained from Hybrid Feature Selection is 82.61% which is higher than the PSO and RR algorithms. Also obtained higher kappa statistic value. While comparing with other measures like MAE, RMSE, RAE, RRAE, TPR, FPR, Precision, Recall and ROC Area, the optimal reduct data set obtained through Hybrid Feature Selection algorithm shows promising results.

Conclusion:-

It is observed that the proposed hybrid methods gives only maximum of 30 attributes, and the existing methods like Relative Reduct and Particle Swarm Optimization gives more than 30 attributes. The error rates like Mean Absolute Error, Root Mean Squared Error, and Relative Absolute Error and Root Relative Absolute Error values are reduced in the proposed hybrid than the existing techniques. True Positive Rate is increased whereas the FPR (False Positive Rate) is reduced in the proposed method. And the values of Precision and Recall are also slightly increased than the existing one. From the table 4, it can be concluded that the proposed hybrid method performs well in all the aspects than the existing methods.

References:-

1. Vajrapu Anusha, Banda Sandhya, "A Learning Based Emotion Classifier with Semantic Text Processing", Advances in Intelligent Systems and Computing, 2015, pp.371-382.
2. Heng Chen, Hai Jin, Feng Zhao, Hanhua Chen, Fei Fang, "A Novel Vector Representation Model for Text Mining Based on Enhancing Features," Journal of Internet Technology, Vol. 16 No. 3, PP. 476-485, 5 2015
3. Verma V.K, Ranjan M, Mishra P, "Text Mining and Information Professionals: Role, issues and Challenges", Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on 6-8 January 2015, pp.133-137.
4. Xiang Ren, Ahmed El-Kishky, Chi Wang and Jiawei Han, "Automatic Entity Recognition and Typing from Massive Text Corpora: A Pharse and Network Mining Appraoch", PMC US National Library of Medicine National Institutes of Health, August 2015, pp.2319-2320.
5. SARVNAZ KARIMI and CHEN WANG and ALEJANDRO METKE-JIMENEZ and RAJ GAIRE and CECILE PARIS, "Text and Data Mining Techniques in Adverse Drug Reaction Detection", ACM Computing Surveys, Vol. 1, No. 1, Article 1, January 2015, pp.1-37.
6. Hsin-Chang Yanga, Chung-Hong Lee, Han-Wei Hsiao, "Incorporating Self-Organizing Map with Text Mining Techniques for Text Hierarchy Generation", Applied Soft Computing, April 2015, pp.1-25.
7. Naw, Naw, and Ei Ei Hlaing. "Relevant words extraction method for recommendation system." Bulletin of Electrical Engineering and Informatics 2.3 (2013): 169-176.
8. Adeva, JJ García, et al. "Automatic text classification to support systematic reviews in medicine." Expert Systems with Applications 41.4 (2014): 1498-1508.
9. Lima, Rinaldo, Bernard Espinasse, and Fred Freitas. "Relation Extraction from Texts with Symbolic Rules Induced by Inductive Logic Programming." Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on. IEEE, 2015.
10. Menaka. S and N. Radha, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Software Engineering, pp.2013.
11. Koplenig, Alexander, et al. "The statistical trade-off between word order and word structure-large-scale evidence for the principle of least effort." arXiv preprint arXiv:1608.03587 (2016).
12. Iqbal, Farkhund, et al. "A unified data mining solution for authorship analysis in anonymous textual communications." Information Sciences 231 (2013): 98-112.
13. Taghandiki, Kazem, Ahmad Zaeri, and Amirreza Shirani, "A Supervised Approach for Automatic Web Documents Topic Extraction Using Well-Known Web Design Features." (2016).
14. obon-Mejia, Diego Alejandro, et al. "A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models." IEEE Transactions on reliability 61.2 (2012): 491-503.
15. Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political Analysis (2013): mps028.
16. Anami, Basavaraj S., Ramesh S. Wadawadagi, and Veerappa B. Pagi. "Machine learning techniques in Web content mining: a comparative analysis." Journal of Information & Knowledge Management 13.01 (2014): 1450005.

17. Xing Zhai, Zhihong Li, Kuo Gao, Youliang Huang, Lin Lin, Le Wang, "Research Status and Trend Analysis of Global Biomedical Text Mining Studies in recent 10 years", Scientometrics, Volume 105, Issue 1, October 2015, pp.509-523.
18. Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S. Kohane, Tianxi Cai, "Toward High-Throughput phenotyping: unbiased Automated Feature Extraction and Selection from Knowledge Source", Journal of the Americal Medical Informatics Association, 2015, pp.993-1000.
19. Basant Agarwal and Namita Mittal, "Sentiment Classification using Rough Set based Hybrid Feature Selection", Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis- Association for Computational Linguistics, pp.115-119, 2013.
20. Girish Chandrashekar, Ferat Sahin, "A Survey on Feature Selection Methods", Computers and Electrical Engineering-Elsevier, pp.16-28, 2014.
21. Jan Platos, Vaclav Snasel, Tomas Jezowicz, Pavel Kromer, Ajith Abraham, "A PSO-Based Document Classification Algorithm accelerated by the CUDA Platform", 2012 IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea.
22. Xiangyang Wang, Jie Yang, Xialong Tens and Weijan Xia, Richard Jension, " Feature selection basedon Rough Set and Particle Swarm Optimization", *Pattern Recognition Letters*, 2007, pp: 459-471.
23. http://en.wikipedia.org/wiki/Rough_set
24. Dataset is collected from KEEL Repository. Dataset Source: http://sci2s.ugr.es/keel/