International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

An Artificial Intelligence Approach for Web Mining Optimization

Swarnima Mishra
Computer Science & Engg.
U.P.T.U., Lucknow (U.P.)
swarnima_mishra@yahoo.com

Dr. S. P. Tripathi
Dept. of Computer Science & Engg.
I. E. T., Lucknow (U.P.)
tripathee_sp@yahoo.co.in

***Abstract*-Due to rapid advances in data generation, availability of automated tools in data collection and continued decline in data storage cost has resulted in high volumes of data. Also, the data is non scalable, highly dimensional, heterogeneous and complex in its nature. This has led to an increasing demand for retrieval of meaning data, in turn leading to the evolution of web mining as a vital research area. In the present work, the application of genetic algorithm framework using Rapid Miner tool in the problem of knowledge discovery in web databases has been explored. The study has focused on genetic algorithms model, leading to attribute reduction for website optimization. Online News Popularity dataset made available from UCI Machine Learning Repository has been used for the analysis. The results obtained clearly showed the superiority of GA based optimised models for web site optimization having a lower RMSE value of 72.443.**

Keywords: Genetic Algorithm, Web Mining, Optimization, Rapid Miner.

1 Introduction

The Web is a massive, varied, dynamic and mostly unstructured data repository, which provides incredible amount of data information, and also increases the complexity of how to deal with the information from the different perceptions of users, Web service providers, business analysts etc. [6].Web mining is divided into three areas: Web content mining (WCM), Web structure mining (WSM), and Web Usage mining (WUM). Web content mining is a process of picking up information from texts, images, audio, video, or structured records such as lists and tables and scripts. Web structure mining is a process of discovering structure information from linkages of web pages (inter page structure/hyper link structure). The web usage or log mining is defined as the process of extracting interesting patterns from the log data. The log data is consists of textual data and is represented in standard format (common log format or extended log format).The main goal of web usage mining is to capture, model and examine the web log data in such a way that it inevitably determines the usage behaviour of web user [6]. In the present work, the application of genetic programming framework in the problem of knowledge discovery in web databases has been explored. The study has focused on genetic algorithms model. Further, the possibility of applying genetic programming in web mining has been surveyed, examining its integration with databases and possibilities of attribute set reduction.

This paper is organized as follows. Section 2 presents related work on the subject. Section 3 describes the dataset used. Section 4 explains the details of the methodology used. Section 5 is about GA based model development, section 6 is results and discussions about the model developed on a real web site structure. Finally in section 7, the conclusion and future works are presented.

2. Related Work

Genetic Algorithms have been applied with success on web mining. In the field of information retrieval for search engines many GA based model have been developed. [7] proposed the Web Structure according to an effectiveness criterion. Showed that researching probabilistic behaviours have permitted quality development of the hyperlink structure, using a Hybrid GA/PSO. [8] dealt with a challenging association rule mining problem of finding frequent itemsets from XML database using proposed GA based method. It was successfully tested for different XML databases. [9] introduced the Advanced Optimal Web Intelligent Model with granular computing nature of Genetic Algorithms. [10] presented the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com. The main purpose is to apply a set of descriptive data mining techniques to obtain rules that allow to help to the webmaster team to improve the design of the website. [11] introduced a genetic algorithm, to be used in a model-driven decision-support system for web-site optimizations.

3. Dataset Used

In the present work the data that is going to be used is the Online News Popularity dataset available from UCI Machine Learning Repository [1]. It has 58 predictive attributes, 2 non-predictive, 1 goal field. In the present work out of 58 attributes, only 10 have been used for model development. The dataset features are given in Table-1 below. The output of the model is the number of shares (shares**)** of the news sites. Data are going to be used for predicting the number of attributes (input variables) using the number of shares of different online news channels as labels (output variables) using GA.

Table 1 : List of Attributes used for model development

Sl. No.	**Attribute Names**	**Variable Type**
1	url: URL of the article	Id
2	n_tokens_title: Number of words in the title	Input
3	n_tokens_content: Number of words in the content	Input
4	n_non_stop_unique_tokens: Rate of unique non-stop words in the content	Input

5	num_hrefs: Number of links	Input
6	num_imgs: Number of Images	Input
7	average_token_length: Average length of the words in the content	Input
8	max_positive_polarity: Max. polarity of positive words	Input
9	abs_title_subjectivity: Absolute subjectivity level	Input
10	Shares: Number of shares (target)	Label

4. Methodology Used

Genetic Algorithms (GAs) are heuristic search algorithm based on the ideas of natural selection and genetic. GA is an approach to inductive learning. GA works in an iterative manner. It uses fitness measure to solve the problem [4]. Standard GA uses genetic operators such selection, crossover and mutation. GA runs to generate solutions for successive generations. Hence the quality of the solutions in successive generations improves [5], the process is terminated when an optimum solution is found.

5. GA Based Model Development

The present study deals with the development of a GA model for the web site optimization, leading to attribute set reduction, based on the number of shares of the most popular online news sites. For this a two stage data analysis has been done on MATLAB [2] and RapidMiner platform. Initially the dataset was preprocessed using Linear Regression Technique. Next, comes the prediction of the reduced number of attributes for website development using GA based modelling in RapidMiner.

5.1 Data Preprocessing

Data per-processing is an important step in data mining. It helps to remove garbage information effect from initial data set. It intends to reduce some noises, incomplete and inconsistent data. Here in order to handle the outliers, during analysis of the initial dataset Robust regression techniques was implemented on the model datasets. For this *robustfit* method was used in MATLAB. The function implements a robust fitting method that is less sensitive than ordinary least squares to large changes in small parts of the data.

5.2 Model Used for Optimization using GA

GA modeling for estimation of the popularity (number of shares) of the most popular online news channels has been carried out in this work using the RapidMiner 5 software. The flowchart for carrying out the attribute set reduction and transformation is given below.

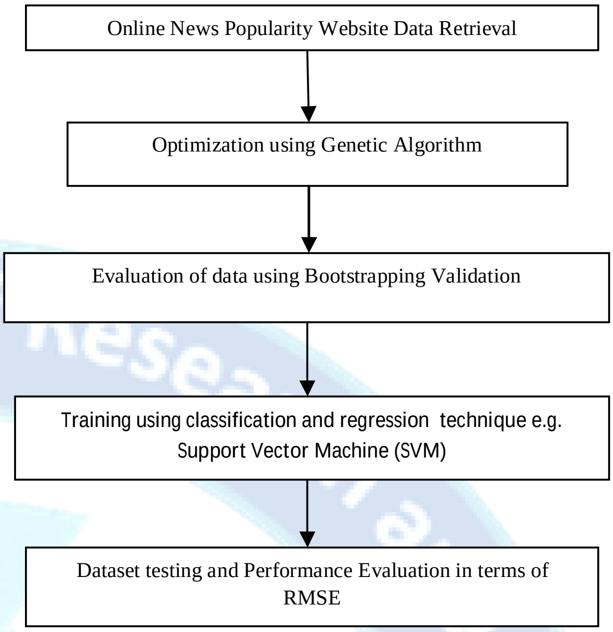

Online News Popularity Website Data Retrieval
Optimization using Genetic Algorithm
Evaluation of data using Bootstrapping Validation
Training using classification and regression technique e.g. Support Vector Machine (SVM)
Dataset testing and Performance Evaluation in terms of RMSE

Fig.1 : A simple flowchart of GA based process implemented

5.3 Modelling Steps Used for Optimization of Datasets

5.3.1 Optimization of datasets using Evolutionary Genetic Algorithm

This technique uses the Evolutionary Genetic Algorithm, wherein it selects the really applicable attributes of the given dataset. Here Genetic Algorithm has been used for attribute selection. Thus the question for the most pertinent features for classification or regression problems, is one of the most important data mining tasks.

5.3.2 Dataset Evaluation using Bootstrapping Validation technique

Bootstrapping sampling is sampling with substitution. Here, at each step all examples have same probability of being selected. Once it has been selected in the sample, it becomes a candidate solution and has further chance of being selected again in the coming steps. Hence a replacement sample can have similar example many times. It can be seen that a sample with replacement is more importantly can be used to generate samples with bigger size than the original example set.

5.3.3 Training Dataset using Support Vector Machine (SVM)

SVM is a non-probabilistic binary linear classifier in the sense that it takes a set of input data and then forecasts which of the two possible classes comprise the inputs, for each given input. An SVM training algorithm makes a model that allocates new examples into one category or the other, for a given training examples, each defined as belonging to one of two categories.

5.3.4 Model Testing and Performance Evaluation

First a model is trained on a dataset ; information related to the dataset is learnt by the model. Then the same model can be applied on another dataset usually for prediction. All needed parameters are stored within the model object. It is required that both datasets should have exactly the same number, order, type and role of attributes.

The output of the model is fed to the performance evaluator, wherein the performance of the developed model is determined in terms of RMSE and MSE.

5.4 Screen Shots of the Development Model using Rapid Miner Tool:

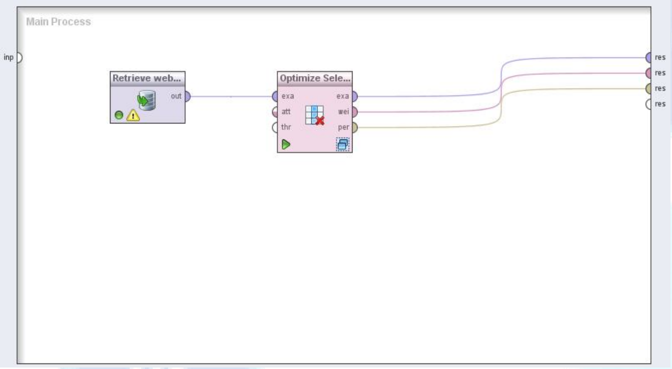


Fig. 2: Screen shot of the Main Process using GA

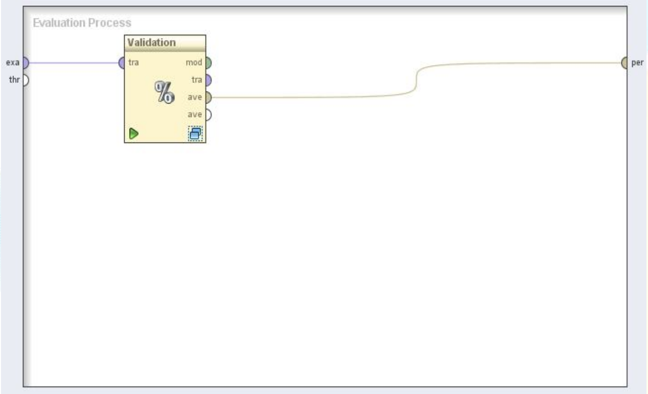


Fig. 3: Screen shot of the Evaluation Process using Bootstrapping Validation Process

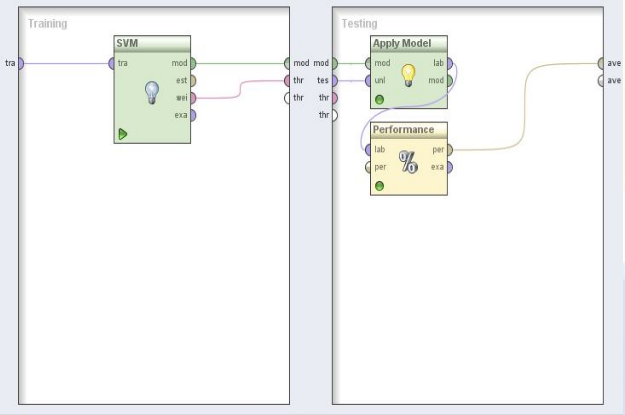


Fig. 4: Screen shot of the Training and Testing Process using SVM

The parameter values used for at different steps for carrying out attribute set reduction and transformation using Evolutionary Genetic Algorithm is given in the tables below.

Table 2: Details of Parameter Values used for Attribute Set Reduction and Transformation

Sl No.	Parameter Used For Attribute Set Reduction and Transformation using Genetic Algorithm	Parameter Values Used
A	**For Evolutionary Genetic Algorithm**	
	Minimum number of attributes	1
	Population size	5
	Maximum number of generations	30
	Maximum fitness	Infinity
	Selection scheme	Tournament
	Tournament size	0.25
	Crossover type	Uniform
	p_initialize	0.5 and 0.75
	P_crossover	0.5 and 0.75
	P_mutation	-1.0, 0.5, 0.75, 1.0
B	**Evaluation Process using Bootstrapping validation**	
	Number of validations	5
	Sample ratio	3, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0
C	**Training dataset using Support Vector Machine**	
	Kernel type	Dot
	Kernel cache	200
	Convergence epsilon	0.001
	Maximum iterations	100000

6. Results and Discussions

As can be seen that the initial dataset attributes were 10 in numbers, with 2 special attributes and 8 regular attributes. These 441 datasets were analysed using RapidMiner tool.

The main process of genetic optimization was begun keeping the minimum number of attributes as one, population size as 5, maximum number of generations as 30, maximum fitness value as infinity, selection scheme as Tournament, tournament size as 0.25, cross over type as Uniform and lastly various permutation and combinations of p crossover, p mutation and p initialize.

During the Evaluation process the number of validations were kept as 5 and the sample ratio was increased from 3.0 to 6.0, with an increment of 0.5. Here, during the training phase, Support Vector Machine (SVM) learning method was used. This learning method can be used for both regression and classification and provides a fast algorithm and good results for many learning tasks. For this kernel type was kept to dot, kernel cache was kept as 200, convergence epsilon as 0.001 and maximum iterations as 100000.

Keeping the above parameters for carrying out the optimization process, it was seen that the best attribute sets were reduced originally from 8 to 5, having the RMSE value of 72.443 (screen shot shown below). Other results obtained from using the various combinations of parameters of GA and Bootstrapping validation techniques are given in the tables 3 and 4 below.

Table 3: Showing the best RMSE values for reduced set of attributes

Sl. No.	Original number of Attributes	Reduced set of attributes	Sample Ratio	Attribute Names	RMSE Values

1	10, 2 special and 8 regular	3	3.0	F, ***G, H***	708.729
2	10	5	3.5	C,D,F,***G,H***	599.437
3	10	5	4.0	A,D,F,***G,H***	577.402
4	10	4	4.5	A,E,F,***H***	443.463
5	10	5	5.0	C,D,F,***G,H***	301.623
6	10	6	5.5	A,B,C,D,***G,H***	234.717
7	10	4	5.5	D,E,***G,H***	282.894
8	10	3	5.5	E,***G,H***	222.483
9	10	3	5.5	E,***G,H***	237.261
10	10	1	***5.5***	***H***	229.104
11	10	6	5.5	A,B,C,D,***G,H***	223.123
12	10	5	5.5	A,B,C,D,F	72.443

Table 4 : Affect of Sample Ratio on RMSE Values

Sample Ratio	RMSE
3	708.729
3.5	599.437
4	577.402
4.5	443.436
5	301.623
5.5	234.717

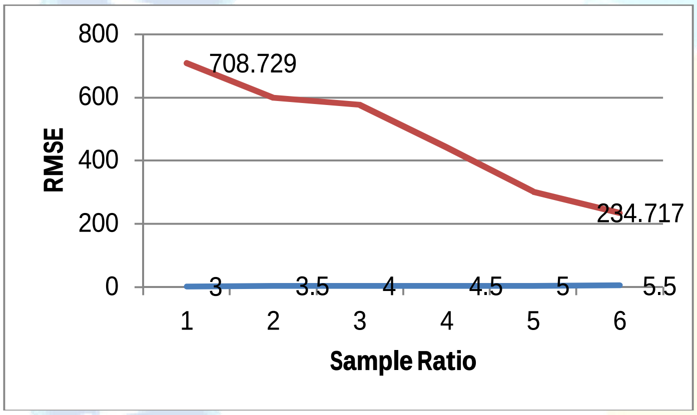
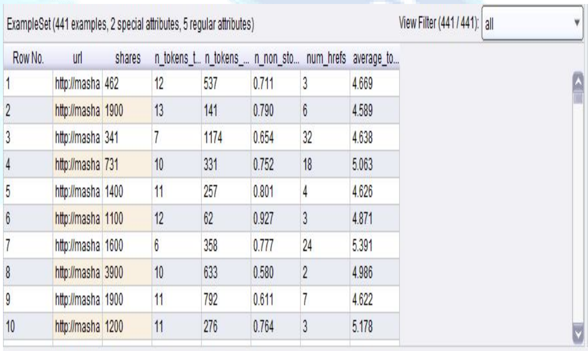


Fig. 5: RMSE Variation in relation to change in Sample Ratio

From the perusal of the data given in table 4, it is clear that sample ratio plays a very important role in the optimization process. As can be interpreted from the table 4, as the sampling ration is increased from 3.0 onwards till 5.5, the RMSE valus for attribuite set reduction also goes on decreasing. This can be very well seen from the analysis of Table 3 and Fig. 5. However it was further seen that increasing the sample ratio from 5.5 to 6.0, the results were unknown. Hence the reduction in sample ration was stopped at 5.5, where the RMSE value was 234.717.

Next, now keeping the sample ratio constant at 5.5, the other genetic algorithm parameters were shuffled. The various permutation and combinations of the probability of attributes, viz, *p_initialize, p_mutation and p_crossover* were used, as depicted in Table 3 below. The best RMSE value of 72.443 was obtained for p mutation value 0.5, p crossover value 0.75 and p initialize value 0.75.

The most striking feature which is noticed is that the best attribute set reduction, havings RMSE value of 72.443 does not have two attributes viz. ***G and H.*** This clearly shows that attributes ***max_positive_polarity*** and ***abs_title_subjectivity*** are the least influencing factors for the best website development. Whereas on the other hand the best online news popularity web site has four attributes A, B, C, D and F, devoid of G and H. The screen shots of the best reduced attributes sets is given in figure 6 below.


ExampleSet (441 examples, 2 special attributes, 5 regular attributes)

Row No.	url	shares	n_tokens_t...	n_tokens_...	n_non_sto...	num_hrefs	average_to...
1	http://masha	462	12	537	0.711	3	4.669
2	http://masha	1900	13	141	0.790	6	4.589
3	http://masha	341	7	1174	0.654	32	4.638
4	http://masha	731	10	331	0.752	18	5.063
5	http://masha	1400	11	257	0.801	4	4.626
6	http://masha	1100	12	62	0.927	3	4.871
7	http://masha	1600	6	358	0.777	24	5.391
8	http://masha	3900	10	633	0.580	2	4.986
9	http://masha	1900	11	792	0.611	7	4.622
10	http://masha	1200	11	276	0.764	3	5.178

Fig. 6: Screen Shot of the RapidMiner Output of the reduced set of attributes

7. Conclusions and Future Work

In the present study, applicability and capability of Genetic Algorithm techniques for application in web mining as an optimization tool for attribute set reduction and transformation for best website development has been investigated. It is seen that GA models are very robust, characterised by fast computation, capable of handling the noisy and approximate data that are typical of data used here for the present study. From the analysis of the results given earlier it is seen that GA has been able to perform well for the selection of the optimal number of attributes using various parameters. Due to the presence of non-linearity in the data, it is an efficient quantitative tool prediction application in web mining. The studies has been carried out using MATLAB and RapidMiner tools.

Future work may include input variables could be extended with more attributes, increasing the number of datasets, application of a hybrid approach using GA in combination with other soft computing techniques.

References:

[1]. https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

[2]. Matlab 2012a: Optimization Toolbox - Product Documentation.

[3]. Sita Gupta, Vinod Todwal, (2012), “Web Data Mining & Applications”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, pp. 20-24.

[4]. Ammar Al-Dallal, et. al. , “Genetic Algorithm Based Mining for HTML Document”.

[5]. *D. Kerana Hanirex and K.P. Kaliyamurthie, (2014),* "Mining Frequent Itemsets Using Genetic Algorithm*"*, Middle-East Journal of Scientific Research 19 (6): 807-810.

[6]. Neetu Anand, et. al., (2012), "Identifying the User Access Pattern in Web Log Data", International Journal of Computer Science and Information Technologies, Vol. 3 (2), 3536-3539

[7]. B. RAJDEEPA, DR. P. SUMATHI, (2014), "WEB STRUCTURE MINING FOR USERS BASED ON A HYBRID GA/PSO APPROACH", Journal of Theoretical and Applied Information Technology, Vol.70 No.3, pp. 573-578.

[8]. Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, (2011), " XML MINING USING GENETIC ALGORITHM", Journal of Global Research in Computer Science, Volume 2, No. 5, pp. 86-90.

[9]. V.V.R. Maheswara Rao, Dr. V. Valli Kumari and Dr. K.V.S.V.N Raju, (2010), "An Advanced Optimal Web Intelligent Model for Mining Web User Usage Behavior using Genetic Algorithm", ACEEE Int. J. on Network Security, Vol. 01, No. 03, Dec 2010, pp. 44-49.

[10]. C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, (2012), "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com", Expert Systems with Applications 39, 11243–11249.

[11]. Arben Asllani, Alireza Lari, (2007), "Using genetic algorithm for dynamic and multiple criteria web-site optimizations", European Journal of Operational Research 176 (2007).