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

Web Site Popularity Estimation using Genetic Algorithm based Modelling

Rudrendra Bahadur Singh
Computer Science & Engg. Dept
BBDNITM, Lucknow (U.P.)
rudra.rathor20@gmail.com

Surya Vikram Singh
Computer Science & Engg. Dept
BBDNITM, Lucknow (U.P.)
m.tech.surya@gmail.com

Himanshu Kumar Shukla
Computer Science & Engg. Dept
BBDNITM, Lucknow (U.P.)
himanshu0590@gmail.com

Abstract- **Due to rapid advances in information technology, availability of automatic equipment in statistics collection and persisted decline in records storage value has ended in high volumes of statistics. Also, the records is non scalable, noticeably dimensional, heterogeneous and complex in its nature. This has brought about an growing call for retrieval of which means statistics, in turn leading to the evolution of web mining as a important studies vicinity. In the prevailing work, the utility of genetic algorithm framework the use of Rapid Miner tool in the problem of expertise discovery in internet databases has been explored. The take a look at has centered on genetic algorithms model, leading to attribute discount for internet site optimization. Online News Popularity dataset made to be had from UCI Machine Learning Repository has been used for the analysis. The effects acquired honestly confirmed the prevalence of GA primarily based optimised fashions for web web page optimization having a lower RMSE value of 72.443.**

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

1 Introduction

The Web is a massive, numerous, dynamic and typically unstructured facts repository, which gives high-quality quantity of records information, and also will increase the complexity of the way to cope with the statistics from the exceptional perceptions of users, Web service providers, enterprise analysts and so forth. [6].Web mining is divided into 3 regions: Web content material mining (WCM), Web shape mining (WSM), and Web Usage mining (WUM). Web content mining is a manner of choosing up information from texts, pics, audio, video, or dependent statistics together with lists and tables and scripts. Web shape mining is a method of coming across shape data from linkages of net pages (inter page structure/hyper link structure). The web usage or log mining is defined because the process of extracting exciting styles from the log information. The log facts is includes textual records and is represented in widespread layout (common log format or prolonged log format).The main aim of net usage mining is to seize, model and take a look at the internet log records in this sort of manner that it necessarily determines the utilization behaviour of net consumer [6]. In the present paintings, the application of genetic programming framework in the hassle of information discovery in web databases has been explored. The have a look at has centered on genetic algorithms version. Further, the possibility of making use of genetic programming in net 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 at the concern. Section 3 describes the dataset used. Section 4 explains the info of the technique used. Section five is set GA based version improvement, phase 6 is outcomes and discussions approximately the model advanced on a actual internet website online shape. Finally in phase 7, the realization and destiny works are provided.

2. Related Work

Genetic Algorithms had been applied with success on net mining. In the sector of statistics retrieval for search engines many GA based version had been advanced. [7] proposed the Web Structure in keeping with an effectiveness criterion. Showed that studying probabilistic behaviours have accepted best improvement of the hyperlink shape, the use of a Hybrid GA/PSO. [8] treated a challenging affiliation rule mining hassle of locating frequent itemsets from XML database the usage of proposed GA based totally method. It turned into efficaciously tested for extraordinary XML databases. [9] added the Advanced Optimal Web Intelligent Model with granular computing nature of Genetic Algorithms. [10] offered the methodology used in an e-trade internet site of extra virgin olive oil sale known as www.OrOliveSur.Com. The predominant motive is to use a fixed of descriptive records mining strategies to attain rules that allow to help to the webmaster team to enhance the design of the internet site. [11] brought a genetic algorithm, to be used in a model-driven choice-guide system for web-website optimizations.

3. Dataset Used

In the prevailing paintings the records that is going for use is the Online News Popularity dataset available from UCI Machine Learning Repository [1]. It has fifty eight predictive attributes, 2 non-predictive, 1 goal discipline. In the present workout of 58 attributes, best 10 were used for version improvement. The dataset functions are given in Table-1 underneath. The output of the version is the number of shares (stocks) of the information websites. Data are going to be used for predicting the range of attributes (enter variables) the usage of the quantity of stocks of various on line 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	Input

	of words in the content	
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 set of rules based totally at the ideas of natural choice and genetic. GA is an technique to inductive studying. GA works in an iterative way. It uses health degree to resolve the hassle [4]. Standard GA uses genetic operators such choice, crossover and mutation. GA runs to generate answers for successive generations. Hence the best of the solutions in successive generations improves [5], the process is terminated while an most reliable answer is observed.

5. GA Based Model Development

The present take a look at offers with the development of a GA version for the internet website online optimization, leading to attribute set discount, based totally on the range of shares of the maximum popular online information web sites. For this a two degree statistics analysis has been executed on MATLAB [2] and RapidMiner platform. Initially the dataset become preprocessed the usage of Linear Regression Technique. Next, comes the prediction of the decreased wide variety of attributes for website development the usage of GA based totally modelling in RapidMiner.

5.1 Data Preprocessing

Data in line with-processing is an vital step in data mining. It helps to remove garbage statistics impact from initial records set. It intends to reduce some noises, incomplete and inconsistent data. Here which will deal with the outliers, at some stage in analysis of the preliminary dataset Robust regression strategies turned into implemented on the model datasets. For this robustfit technique become utilized in MATLAB. The feature implements a robust fitting approach this is much less sensitive than regular least squares to massive adjustments in small parts of the statistics.

5.2 Model Used for Optimization the use of GA

GA modeling for estimation of the recognition (quantity of stocks) of the most popular online information channels has been performed on this paintings the use of the RapidMiner five software. The flowchart for carrying out the characteristic set discount and transformation is given beneath.

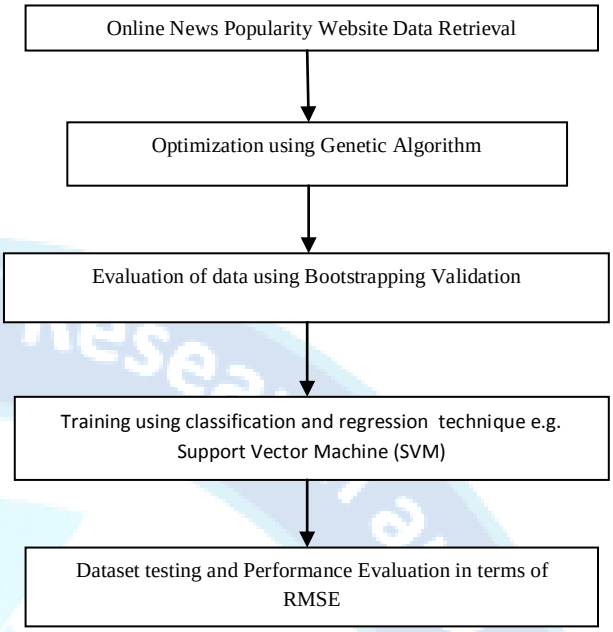

Fig.1 : A simple flowchart of GA based process implemented

5.3 Modelling Steps Used for Optimization of Datasets

5.3.1 Optimization of datasets using Evolutionary Genetic Algorithm

This method uses the Evolutionary Genetic Algorithm, in which it selects the virtually relevant attributes of the given dataset. Here Genetic Algorithm has been used for characteristic selection. Thus the query for the maximum pertinent features for category or regression troubles, is one of the maximum essential data mining obligations.

5.3.2 Dataset Evaluation using Bootstrapping Validation method

Bootstrapping sampling is sampling with substitution. Here, at every step all examples have equal chance of being selected. Once it has been decided on inside the sample, it turns into a candidate solution and has similarly danger of being selected again in the coming steps. Hence a replacement pattern will have comparable instance normally. It can be seen that a sample with substitute is greater importantly may be used to generate samples with bigger size than the authentic example set.

5.3.3 Training Dataset the usage of Support Vector Machine (SVM)

SVM is a non-probabilistic binary linear classifier within the sense that it takes a hard and fast of input statistics after which forecasts which of the 2 viable instructions include the inputs, for every given enter. An SVM schooling algorithm makes a version that allocates new examples into one category or the opposite, for a given training examples, each described as belonging to certainly one of two classes.

5.3.4 Model Testing and Performance Evaluation

First a model is skilled on a dataset ; statistics associated with the dataset is learnt by means of the model. Then the identical model may be applied on every other dataset usually for prediction. All needed parameters are stored within the version item. It is needed that both datasets need

to have precisely the identical number, order, kind and role of attributes.
The output of the version is fed to the overall performance evaluator, in which the performance of the developed model is decided in phrases of RMSE and MSE.

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

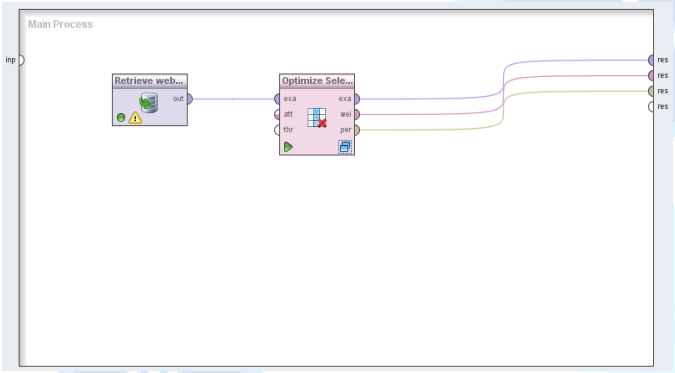

Fig. 2: Screen shot of the Main Process using GA

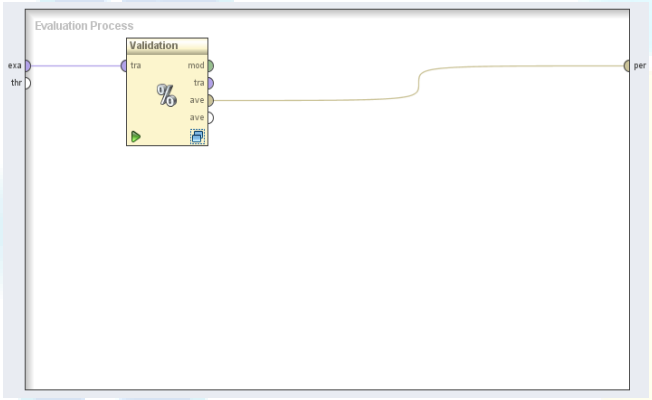

Fig. 3: Screen shot of the Evaluation Process using Bootstrapping Validation Process

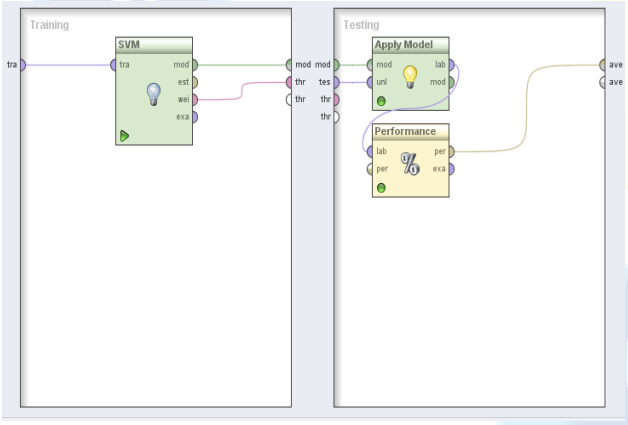

Fig. 4: Screen shot of the Training and Testing Process using SVM

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 preliminary dataset attributes have been 10 in numbers, with 2 unique attributes and 8 ordinary attributes. These 441 datasets were analysed using RapidMiner device.
The essential technique of genetic optimization turned into all started maintaining the minimal quantity of attributes as one, populace length as five, most number of generations as 30, most fitness value as infinity, choice scheme as Tournament, tournament size as zero.25, pass over type as Uniform and lastly numerous permutation and combinations of p crossover, p mutation and p initialize.
During the Evaluation manner the range of validations have been kept as 5 and the sample ratio become multiplied from 3.0 to 6.0, with an increment of 0.5. Here, throughout the training segment, Support Vector Machine (SVM) studying technique become used. This mastering approach may be used for each regression and class and offers a fast set of rules and excellent outcomes for lots mastering duties. For this kernel kind was stored to dot, kernel cache become kept as two hundred, convergence epsilon as zero.001 and maximum iterations as one hundred thousand.
Keeping the above parameters for sporting out the optimization method, it became seen that the fine characteristic sets were reduced originally from eight to five, having the RMSE price of 72.443 (screen shot shown underneath). Other effects acquired from the usage of the diverse combos of parameters of GA and Bootstrapping validation techniques are given inside the tables 3 and 4 under.

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

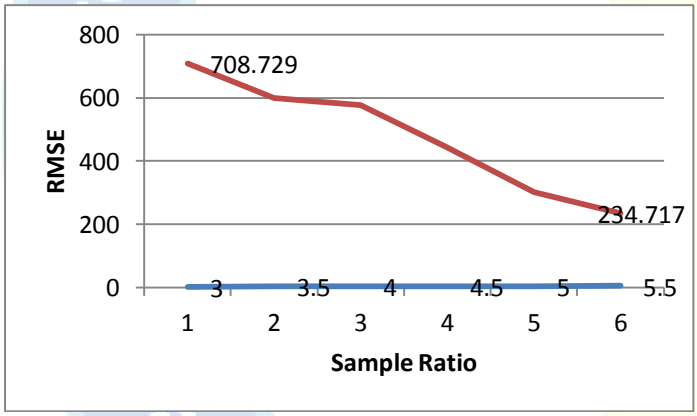

Fig. 5: RMSE Variation in relation to change in Sample Ratio

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.

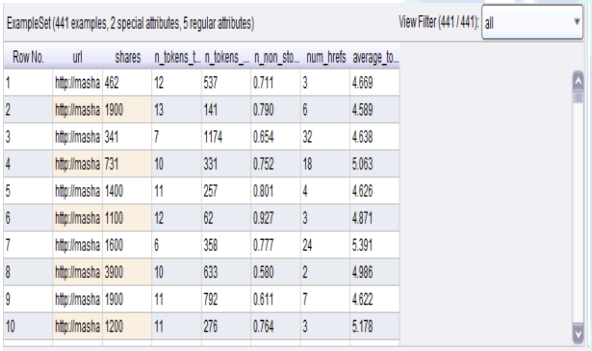

Fig. 6: Screen Shot of the RapidMiner Output of the reduced set of attributes

7. Conclusions and Future Work

In the prevailing study, applicability and capability of Genetic Algorithm strategies for application in web mining as an optimization tool for attribute set discount and transformation for excellent internet site improvement has been investigated. It is seen that GA models are very sturdy, characterized by way of rapid computation, able to handling the noisy and approximate facts which can be normal of statistics used right here for the present have a look at. From the evaluation of the effects given earlier it's far visible that GA has been able to carry out nicely for the choice of the ultimate wide variety of attributes using numerous parameters. Due to the presence of non-linearity inside the information, it's far an efficient quantitative device prediction utility in net mining. The studies has been achieved using MATLAB and RapidMiner gear.

Future paintings may consist of enter variables could be extended with more attributes, growing the variety of datasets, application of a hybrid technique using GA in mixture with other smooth 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).