Optimization-Based Effective Feature Set Selection in Big Data

Chapter · May 2020

DOI: 10.1007/978-981-15-3284-9_7

Optimization-Based Effective Feature Set Selection in Big Data

J. S. T. M. Poovarasi, Sujatha Srinivasan, and G. Suseendran

Abstract Of late, the data mining has appeared on the arena as an ideal form of knowledge discovery crucial for the purpose of providing appropriate solutions to an assortment of issues in a specified sphere. In this regard, the classification represents an effective method deployed with a view to locating several categories of anonymous data. Further, the feature selection has significantly showcased its supreme efficiency in a host of applications by effectively ushering in easier and more all-inclusive remodel, augmenting the learning performance, and organizing fresh and comprehensible data. However, of late, certain severe stumbling blocks have cropped up in the arena of feature selection, in the form of certain distinctive traits of significant of big data, like the data velocity and data variety. In the document, a sincere effort is made to successfully address the prospective problems encountered by the feature selection in respect of big data analytics. Various tests conducted have upheld the fact that the oppositional grasshopper techniques are endowed with the acumen of effectively extracting the requisite features so as to achieve the preferred outcome Further, enthusing experimental outcomes have revealed the fact only a trivial number of hidden neurons are necessary for the purpose of the feature selection to effectively appraise the quality of an individual, which represents a chosen subset of features.

Keywords Classification · Optimization · Oppositional grasshopper

J. S. T. M. Poovarasi (✉)
Research Scholar, School of Computing Science, Vels Institute of Science,
Technology and Advanced Studies, (VISTAS), Chennai, Tamil Nadu, India

S. Srinivasan
Associate Professor, Department of Computer Science and Applications,
SRM Institute for Training and Development, Chennai, Tamil Nadu, India

G. Suseendran
Assistant Professor, School of Computing Science, Vels Institute of Science,
Technology & Advanced Studies, (VISTAS), Chennai, Tamil Nadu, India

S.-L. Peng et al. (eds.), *Intelligent Computing and Innovation on Data Science*,
Lecture Notes in Networks and Systems 118,
https://doi.org/10.1007/978-981-15-3284-9_7

1 Introduction

Incidentally, the big data, in quintessence, represents the group of incredibly massive information sets with wide variety of categories, thereby making it exceedingly hard to process them by employing the high-tech data processing techniques or time-honored data processing platforms. No wonder, the big data has affected a sea change in our traditional styles of businesses, administrations and experimentations. The information exhaustive science, especially in information exhaustive calculating, has appeared on the stage which is dedicated for the launch of the requisite devices to outwit the hassles encountered by the big data. The big data kicks off with colossal, diverse and independent distributed resources and controlled decentralization are effectively processed to testing the intricate and budding bonds among the data. The relative traits emerge as severe hassles in the process of locating fruitful information from the big data [1]. Moreover, the data finds itself saved in the disseminated system files like the MapReduce/Hadoop. Hence, it is all the more essential to storage, query and communication troubles. In certain instances, the private constraints effectively withhold entire information set, permits only the preprocessed information is communicated by means of cautiously devised interfaces. On account of their probable incongruent origins, the big data sets are generally found to be imperfect, with a large segment being misplaced. In fact, the mammoth quantum of data invariably possess tainted measurements, communication faults, in addition to being prone to the severe cyber assaults, especially when the overheads relating to purchase and transport per entry are reduced to the least [2]. The big data, in essence, represents the extensively employed term indicating huge collection of datasets which are so highly complicated that it is very hard to process them by employing the time-honored data processing applications. The various types of challenges in this regard are such as the assessment, pattern identification, visualization and the likes. Usually, the big data assessment is effectively carried out in various spheres such as the cloud environment, network simulation and forecast and so on [3]. The distinguishing procedure of a pattern identification method basically decreases the dimensionality of input data into the different classes. As a matter of fact, the dimensionality decrease is extensively observed unreservedly in the whole modules of the identification mechanism such as the preprocessing, feature extraction and classification [4]. Now a days, the analysis of big data is slowly emerging as key for creative values of applications and modern enterprises, these are arranged as the accumulate direct customer reaction data from the business processes internally [5]. In fact, the big data invariably characterizes the typical dominion of issues and methods employed for the application domains which collect and preserve gigantic quantity of unrefined data for the domain-specific data assessment. The current data-intensive methods and the improved computational and data storage resources have played a significant part in the advancement of the big data science [6].

A lion's share of the reduction dimensionality techniques has concentrated on the features which operate with the maximum significance to the target class [7]. A lot of investigations have been conducted on the dimensionality decline in the region of the synchrophasor data. Predominantly, the online dimensionality diminution aims at the extraction of correlations among the synchrophasor measurements, like the voltage, current, frequency and so on [8]. An extraction of features, in turn, represents a vital technique dedicated for the purpose of extracting fruitful data hiding within the electromyography (EMG) signal, ignoring redundant part and interventions [9]. The big data applications are extensively and fruitfully employed in various scientific controllers like parallel complicated and inter-controlled scientific investigation [10].

2 Problem Definition

Dimensionality decrease is invariably targeted at the adaptation of high-dimensional data into an aligned low-dimensional illustration. It effectively executes the function of significantly scaling down the computational intricacy and improves the statistical ill-conditioning by way of eliminating the superfluous traits which is likely to weaken the classification efficiency. In certain applications like detection of optic device, recognition, bioinformatics, and data mining and high information dimensionality put several roadblocks in the path of the vigorous and precise identification. Moreover, the organization and scrutiny of medical big data are beset with a host of varied problems in regard to their structure, storage and analysis. It is, indeed, a Herculean task to accumulate and process the colossal quantity of data generated in the big data. In comparison to the parallel problems encountered by the big data, inadequate consideration it paid to the sampling issue. In view of constraints such the space and time, it has become an extremely hard task to process the whole big data set simultaneously. The feature selection issue involves the decrease of the number of variables in the input set simultaneously generating the identical output. It is also likely that the values detachable from the input set do not hold fruitful data.

3 Proposed Methodology

The current document makes an earnest effort to conduct a distinctive appraisal of a host of diverse feature selection methods and classification approaches extensively employed for the purpose of mining. The detection of features plays major part on the course of extraction in fruitful information from a dataset. In fact, the distinct features are likely to be interrelated and hence have to be scrutinized in groups instead of examining them individually, which make feature selection procedure further difficult. In the document, the corresponding goal of the selecting feature for big data analytics is envisaged. It is illustrated by means of test conducted that the oppositional grasshopper algorithm is well endowed with the requisite skills to effectively extract the relevant features essential.

4 Hadoop MapReduce Frame Work

The Hadoop MapReduce, in quintessence, represents a software framework which invariably allows the distributed processing of gigantic quantity of data such as the dataset for multi-terabyte in high number of service nodes hardware in a reliable, fault-tolerant basis. In fact, the MapReduce job normally splits into the input dataset into several autonomous structures that are carried out via the map functions in an entirely parallel method. The outputs of the map functions are duly arranged by the framework, for furnishing them as the reduced tasks from the input. Normally, inputs and the outputs of the task are duly saved based on the file system. The novel technique is competent to successfully address the computation issue by means functions of two distinct like map and reduce. Basically, the map reduction technique duly empowers users to write map and diminish the elements with the help of the functional-style code. At last, the relative elements are scheduled by means of the MapReduce system to the scattered assets for implementation in the course of managing a large number of thorny issues like the network communication, parallelization and fault tolerance. First and foremost, the input dataset is duly furnished as the input to the mapped. It is effectively used for the parallel processing of data with elevated speed regardless of the dimension of the data. With the result, it offers a helping hand to significantly scale down the run-time. By means of effective application of the mapper, the big data is duly grouped in a number of clusters. The functional stream of the MapReduce technique contains an input dataset, which, in turn, is categorized into a large number of data components, each of which is effectively administered by the map task in the map segment. Finally, it is joined to the reduce task in the reduce segment to create the eventual consequence. In the MapReduce function, the big computations are easily parallelized and re-accomplishment of futile tasks is deemed as the key technique for the error acceptance. All of these are all represented as the principal compensation of the MapReduce. Mapper is effectively employed with each and every input key-value couple to generate an arbitrary quantity of intermediate key-value couples. The characteristic declaration is well illustrated in Expression (1) shown below.

$$\text{map(in Key, in Value)} \rightarrow \text{list(intermediate key, intermediate Value)} \tag{1}$$

Reducer, it is utilized with each and every value connected by the identical intermediate key with the intention of generating the output key-value couples. The following Expression (2) effectively exhibits the distinctive declaration.

$$\text{reduce(intermediate Key, list(intermediate Value))} \rightarrow \text{list(out Key, out Value)} \tag{2}$$

4.1 Feature Selection

The feature selection (FS) has, of late, emerged as daunting function devoted to the task of diminishing the number of features by way of the eradication of the immaterial, superfluous and noisy data, simultaneously upholding a desirable level of classification precision. In fact, it may be deemed as an optimization issue. In the back of the inherent intricacy of the corresponding problem and amidst a flood of local solutions, the stochastic optimization techniques emerge as the ideal candidates with the necessary acumen to overwhelm the relative issue. As a decisive endeavor, the modified oppositional grasshopper optimization algorithm (MOGOA) inelegantly launched in the document which is effectively worked topic the feature subset for the purpose of types in architecture.

4.2 Modified Oppositional Grasshopper Optimization Algorithm

Here, an inefficient feature selection procedure assisted by the modified oppositional grasshopper optimization technique (MOGHO) is proficiently carried out. For the purpose, an adaptive neural network approach is introduced for precise feature selection process as a fitness function for enhanced precision. Incidentally, the grasshopper represents one of the insects in our biodiversity. Extensively present in the environment, the grasshoppers unite with one among the major swarm of the entire creatures. As the dimension of the swarm is continental in scale, it has become a nightmare for the agriculturists. The nature-motivated techniques rationally classify the search process into two distinct behaviors such as the exploration and the exploitation. In the exploration phase, the analyzing agents are motivated to travel by making hasty moves to longer distances, whereas their travel is limited locally with slow and small steps in the exploitation phase. The target seeking by means of these two functions are carried out by the grasshoppers. It is possible to devise an innovative nature-motivated technique by way of calculation mathematically with the help of the novel activity model.

Step 1: The arithmetical model is effectively worked to replicate the swarming conduct the grasshoppers as illustrated in the following Eq. (1).

$$L_i = C_i + M_i + W_i \tag{1}$$

where L_i characterizes the location of the ith grasshopper, C_i indicates the common interface, M_i signifies the magnitude energy and W_i symbolizes the wind speed convection.

Step 2: With a view, the conventional modernize grasshopper technique; the oppositional method is elegantly brought into limelight. Based on the learning opposition (OBL) propounded through Tizhoosh, the recent agent and their opposite agents are envisaged concurrently so as to realize superior similarity for the recent agent solution. This is taken for granted opposite agent result holds the superior prospect of being near the global optimal result rather than the random agent result. The opposite variance blocks positions (OPb_t) are totally calculated busing components of P_m as illustrated in Eq. (2) below.

$$\mathrm{OPb}_t = \left[\mathrm{opb}_t^1, \mathrm{opb}_t^2, \ldots \mathrm{opb}_t^d\right] \tag{2}$$

Let $\mathrm{OPb}_t = \mathrm{Low}b_t + \mathrm{Up}b_t - Pb_t$ with $\mathrm{OPb}_t \in \left[\mathrm{Low}b_t, \mathrm{Up}b_t\right]$ represents the location of tth low variance blocks OP_t in the dth dimension of oppositional blocks.
Step 3: It is possible for modernize Eq. (1) so as to provide the arbitrary conduct as to $\mathrm{Pos}_i = q_1 \mathrm{Soc}_i + q_2\ \mathrm{Foc}_i + q_3 W\mathrm{in}_{i,}$ where q_1, q_2, and q_3 duly represent the arbitrary numbers in [0, 1].

$$\mathrm{Soc}_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathrm{soc}\left(dl_{ij}\right) \overset{\wedge}{dl}_{ij} \tag{3}$$

where dl_{ij} indicates the distance among the ith and the jth grasshopper.
Step 4: The s-social forces. It is evaluated by means of the following equation.

$$\mathrm{Soc}(r) = \mathrm{Aeo.e}^{\frac{-r}{\int}} - \mathrm{e}^{-r} \tag{4}$$

where W denotes the attraction intensity, $\int$ represents the scale attractive length. The s duly exhibits the attract way it influences the social interaction such as the ion and oppositional grasshoppers.
Step 5: The function in this interval and F equipment in *equation* is effectively evaluated as per equation shown.

$$\mathrm{Foc}_i = \mathrm{goe.}\, \overset{\wedge}{\mathrm{e}}_g \tag{5}$$

where *goe.* symbolizes the gravitational constant and $\overset{\wedge}{\mathrm{e}}_g$ establishes union vector through the center of earth.
Step 6: The *function W* in Eq. (1) is effectively estimated by means of the following equation.

$$W\mathrm{in}_i = \mathrm{uoe.}\, \overset{\wedge}{\mathrm{e}}_v \tag{6}$$

A constant drift is denoted as c and a unity vector is denoted as $\overset{\wedge}{\mathrm{e}}_v$ toward the earth.
Step 7: A nymph grasshopper does not have any wings; hence, their functions are vastly associated through the direction of wind. By way of function S, F and W in Eq. (1), the equation may be improved as per the following Eq. (7).

$$\mathrm{Pos}_i = \sum_{\substack{j=1\\ j\neq i}}^{N} \mathrm{soc}\left(\left|\mathrm{pos}_j - \mathrm{pos}_i\right|\right)\frac{\mathrm{pos}_{j\_}\mathrm{pos}_i}{\mathrm{doc}_{ij}} - \mathrm{goe}.\,\overset{\wedge}{\mathrm{e}}_g + \mathrm{coe}.\,\overset{\wedge}{\mathrm{e}}_v \tag{7}$$

where s and $\mathrm{Soc}(q) = \mathrm{Aoe.e}^{\frac{-r}{f}} - \mathrm{e}^{-r}$, N characterizes the various location should not fall below a certain threshold. Nevertheless, function (8) can be profitably deployed for the replication of interaction among the grasshoppers.

$$\mathrm{Poc}_i = \left(\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{\mathrm{uoe}.b_d - lb_d}{2}\mathrm{soc}\left(\left|\mathrm{poc}_j^d - \mathrm{poc}_i^d\right|\right)\frac{\mathrm{poc}_j - \mathrm{poc}_i}{\mathrm{doc}_{ij}}\right) + \overset{\wedge}{T_d} \tag{8}$$

where $\mathrm{uoe}.b_d$ is indicates Dth upper bound, lb_d indicates the Dth lower bound.$\mathrm{Soc}(q) = \mathrm{Aoe.e}^{\frac{-r}{f}} - \mathrm{e}^{-r}$, $\overset{\wedge}{T_d}$ implies the value of the Dth dimension in the target and uoe. Corresponds to a reducing coefficient to shrink the zones like comfort, repulsion and attraction. It is also presumed that the A component representing the wind direction is constantly in the direction of the target $\overset{\wedge}{T_d}$. The subsequent position of the grasshopper is estimated taking into consideration its current position as illustrated in Eq (8). Further, the status of the entire grasshoppers is envisioned so as to arrive at the search agent's location over the target.

$$C = c\,\mathrm{high} - \frac{c\,\mathrm{low} - c\,\mathrm{high}}{L} \tag{9}$$

where c high indicates the highest value, c low implies the lowest value 1 illustrates the recent testing and L represents the highest number of iterations. The position of the best goal estimated till now is modernized in every testing's. Further, the factor c is evaluated and applied in Eq. (9). The updating of location is effectively carried out by the testing till an end criterion is met with. Function place and fitness of the best target is, at last, back to the best. Above-mentioned replications and debates reiterate the supreme efficiency of the MOGOA technique in arriving at the global optimum in an analyzing area.

The artificial neural systems, in essence, characterize a maximized computational method entrusted with the function of the replication of the neural configuration and functioning of the human cerebrum. It consists of an interconnected framework of deceivingly delivered neurons which functions as the media for the data exchange. The datasets, in turn, are taken to determine the movement of the input constraints. As a rule, the ANN is founded on diverse optimizations of the weights. In the corresponding numerical expression, MOGHO approach is duly followed with the intention of realizing the superior precision and classification outcomes.

5 Result and Discussion

A roughly estimated dimension is elegantly employed to evaluate the effectiveness of the suggested technique. It is invariably home to a set of technique which follows universal basic estimation methods with the dimensions for evaluation such as the precision, recall and *f*-measures.

For the schema aligning procedure, the precision can be defined as the fraction of derived matches of schema attributes relevant to the schema of table instances as illustrated in the following relation

$$\text{Precision, } P = \frac{|(\text{relevant match}) \cap (\text{derived match})|}{|(\text{relevant match})|}$$

For the schema aligning procedure, the recall may be characterized as the fraction of relevant matches derived to the schema of table instances, as shown below.

$$\text{Recall, } R = \frac{|(\text{relevant match}) \cap (\text{derived match})|}{|(\text{derived match})|}$$

The accuracy of the novel technique is represented by the fraction of the sum of TP and TN to the sum of TN + TP + FN + FP as shown below.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{(\text{TN} + \text{TP} + \text{FN} + \text{FP})}$$

See Figs. 1, 2 and Tables 1, 2.

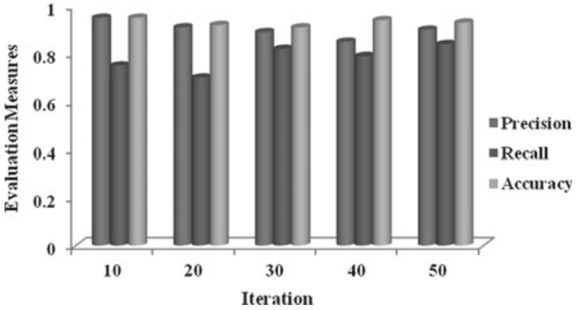

Fig. 1 Graphical representation of our proposed research evaluation measures

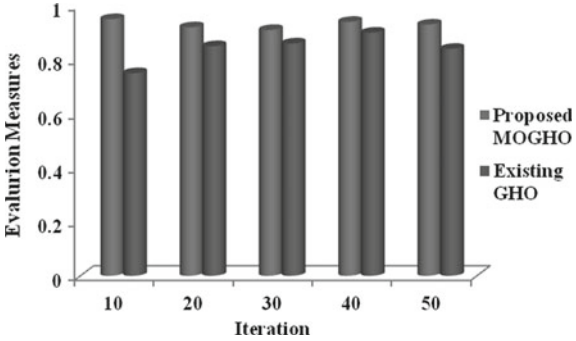

Fig. 2 Graphical representation of proposed and existing accuracy measures

Table 1 Evaluation measures for our proposed research

Iteration	Precision	Recall	Accuracy
10	0.95	0.75	0.95
20	0.91	0.70	0.92
30	0.89	0.82	0.91
40	0.85	0.79	0.94
50	0.90	0.84	0.93

Table 2 Comparison of our presented and existed accuracy measures

Iteration	Presented MOGHO	Existed GHO
10	0.95	0.75
20	0.92	0.85
30	0.91	0.86
40	0.94	0.9
50	0.93	0.84

6 Conclusion

The extensive employment of the big data frameworks to accumulate, process and evaluate data has drastically revolutionized the scenario of the knowledge discovery from data, particularly, the procedures intended for the data preprocessing. In this regard, the feature selection effectively executes its function of lessening certain mapping and classification issues by means of scaling down the number of features to be examined. The new-fangled technique pays scant attention to constraints such as the significance or redundancy of the features, but assigns the relevant task to the artificial neural network, thanks to the exceptional skills exhibited by the latter in the matter of identifying hidden patterns even in the backdrop of noisy scenarios. It is also established without an iota of doubt the genetic algorithm can be effectively employed for the purpose of assisting the search for the relevant features capable of yielding the preferred outcomes. It is hoped that the upcoming researchers, practitioners and data scientists would work hand in hand with the ultimate aim of ensuring the long-term triumph of the big data preprocessing and make a joint move toward the unexplored horizons to quench their thirst.

References

1. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
2. Slavakis K, Giannakis GB, Mateos G (2014) Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge. IEEE Signal Process Mag 31(5):18–31
3. Anjaria M, Guddeti RMR (2014) Influence factor based opinion mining of Twitter data using supervised learning. In process of Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp 1–8
4. Jiang Xudong (2011) Linear subspace learning-based dimensionality reduction. IEEE Signal Process Mag 28(2):16–26
5. ur Rehman MH, Chang V, Batool A, Wah TY (2016) Big data reduction framework for value creation in sustainable enterprises. Int J Inf Manag 36(6):917–928
6. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1
7. Zou Q, Zeng J, Cao L, Ji R (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 173:346–354
8. Diamantoulakis PD, Kapinas VM, Karagiannidis GK (2015) Big data analytics for dynamic energy management in smart grids. Big Data Res 2(3):94–101
9. Phinyomark A, Phukpattaranont P, Limsakul C (2012) Feature reduction and selection for EMG signal classification. Exp Syst Appl 39(8):7420–7431
10. Chen CLP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci 275:314–347