International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Multi-Objective Unsupervised Feature Selection using Genetic Algorithm

Rizwan Ahmed Khan[1], Indu Mandwi[2]

[1]Department of CSE, G.H. Raisoni COE, Nagpur

[2]Assistant Professor, Department of CSE, G.H. Raisoni COE, Nagpur

Abstract: ***Data mining is related to large number of databases. Dealing with such large number of datasets may create some obstacles. Such problems can be avoided by using feature selection Technique. Feature selection Technique is a method which selects an optimal subset from original feature set. The implementation is done by removing repetitive features. The underlying structure has been neglected by the previous feature selection method and it determines the feature separately. The group feature selection method for the group structure may be formulated. It performs the task for filtering purpose for group structure technique. Group feature selection improves accuracy and may achieve relatively better classification performance.***

Keywords: Genetic Algorithm, Supervised Feature Selection, Optimization, classification, K-Nearest Neighbor

1. Introduction

Searching invisible information and pattern from large database is the work of data mining. High dimensionality is become a curse for data mining which generates problem while training the data. The curse of dimensionality can be reduced by using feature selection Technique. works on regression problem to select a group of feature to search the subset of the important factor that achieves efficient prediction [3], The step of Finding an optimal variable subset from original feature set is a feature selection Technique. The application in which there are large numbers of variable the feature selection is enforced to reduce the variable. The aim of feature selection is to find an appropriate feature that is required for target output. It removes the irrelevant and redundant feature from original feature sets. As showed by our broad test think about, the proposed structure accomplishes unrivaled component determination execution and alluring properties [16], Relevant feature provide useful information and redundant feature are those that are not useful than the selected features. So, feature selection is an important phase in efficient learning of large multi-feature data sets.

Both SVM and K-NN classifiers give comparative classification exactness to the baselines (i.e., those without information pre-preparing) [14], There is some potential advantage of feature selection. It enhances data visualization, increases data predictability and understanding. It also helps to minimize the measurement and storage requirement, reduces training and execution time. It is basically observed as being the problem of searching for an optimal feature.

The proposed a feature selection technique where feature appear in a dynamic way [1], Feature selection enhances the learning efficiency, increases predictive accuracy and reducing complexity of learned result. The feature selection algorithm generates an output as a subset of feature or by measuring their utility of feature with weights. We perform broad exploratory examination of our calculation and different strategies utilizing three unique classifiers [18], The idea of features in feature selection can be in various forms such as filtering, consistency, dependency, information and learning model which is generally used in wrapper model.

They execute the feature selection both at the separate level and group level feature selection [2], The feature selection method is mainly classified into three classes based on their label information used most commonly used method label variable. In supervised feature selection, there is difficulty in acquiring the labelled data., make use of label information and multiple structure related to labelled data and unlabelled data.

2. Related Work

This section describes earlier research on feature selection method and to determine the relevant features from groups. The section helps us to understand study design, measurement and data analysis methods and get details of the various methods of group feature selection.

Z.Q. Zhao [1] proposed a feature selection technique where feature appear in a dynamic way and it consists of two steps feature selection. In the first step, it performs the feature selection within the groups to select a selective feature in each group, and in second step it performs the feature selection between the groups by using a LASSO which tends to select an optimal subset of features.

Haiguang Li, [2] focus on the problem where feature contains certain group structure in streaming feature and assume where not all the features are given in advance for feature selection. They execute the feature selection both at the separate level and group level feature selection. This can pick the feature from an important group and pick the feature either at separate feature level or group feature level or both.

M. Yuan and Y. Lin [3] works on regression problem to select a group of feature to search the subset of the important factor that achieves efficient prediction. It considers the group lasso i.e., an expansion of lasso. They correlate the

LARS algorithm and non-negative garrote algorithm. The regression method is used to select a single variable outperformed than the traditional backwards elimination method.

H. Yang [4] Have built an online group lasso algorithm to find important exploratory factor or variable in group manner and displays the limitations of traditional batch mode group lasso algorithm. where data are given in prior, and can manage the data which have several hundred or thousand instances this limitation of batch mode of group LASSO with sparsity by choosing a feature in group level and provide a closed –form solution for group lasso with L1 regularization

Lei Yua [5] has given the overlapping group lasso which is an expansion of lasso, it makes use of L1 norm regularisation and L2-norm for group features for the one upon the other group by using Accelerated gradient descent (AGD) Technique. a less cost processing procedure is built to identify and to erase the zero groups in proximal operation.

S. Xiang [6] analyse the non-convex approach of sparse group features selection for achieving the underlying group structure one by one performs feature selection together. The sparse group model tends to pick the feature both at the individual level and group level. The author executes sparse group feature selection by using constrained non-convex method to optimise L0-model of regularisation norm which represent an efficient optimisation algorithm such as Accelerated gradient method and demonstrate the comparison between convex and non-convex approaches.

M. Ramaswami, et al. [7] As the element determination impacts the prescient precision of any execution show, it is fundamental to contemplate extravagantly the adequacy of understudy execution display regarding highlight choice systems. In this association, the present review is committed not exclusively to explore the most significant subset highlights with least cardinality for accomplishing high prescient execution by embracing different separated element determination procedures in information mining additionally to assess the decency of subsets with various cardinalities and the nature of six sifted include choice calculations as far as F-measure esteem and Receiver Operating Characteristics (ROC) esteem, produced by the Naïve Bayes calculation as benchmark classifier strategy. The relative review did by us on six channel highlight segment calculations uncovers the best technique, and in addition ideal dimensionality of the component subset. Benchmarking of channel highlight determination technique is hence done by sending distinctive classifier models. The aftereffect of the present review successfully underpins the verifiable truth of increment in the prescient exactness with the presence of least number of elements.

Sunita Beniwal, et al. [8] Information mining is a type of learning revelation fundamental for taking care of issues in a particular space. Characterization is a strategy utilized for finding classes of obscure information. Different strategies for grouping exists like bayesian, choice trees, manage based, neural systems and so on. Before applying any mining method, immaterial credits should be sifted. Sifting is done utilizing diverse component determination procedures like wrapper, channel, inserted system.

Huan Liu, et al. [9] The fast progress of PC innovations in information handling, gathering, and capacity has given unparalleled chances to extend abilities underway, administrations, interchanges, and research. Be that as it may, enormous amounts of high-dimensional information restore the difficulties to the best in class information mining systems. Highlight determination is an effective procedure for measurement diminishment and a basic stride in fruitful information mining applications. It is an exploration zone of incredible handy significance and has been created and advanced to answer the difficulties because of information of progressively high dimensionality. Its direct benefits include: building easier and more conceivable models, enhancing information mining execution, and get ready, clean, and comprehend information. We first briefly present the key segments of highlight choice, and audit its improvements with the development of information mining. We then outline FSDM and the papers of FSDM10, which exhibits of a dynamic research field of some contemporary interests, new applications, and progressing research efforts. We then look at early requests in information concentrated applications and distinguish some potential lines of research that require multidisciplinary efforts.

Jinjie Huang, et al. [10] In this review, a mixture hereditary calculation is received to discover a subset of components that are most important to the arrangement undertaking. Two phases of enhancement are included. The external enhancement organize finishes the worldwide scan for the best subset of elements in a wrapper path, in which the common data between the prescient marks of a prepared classifier and the genuine classes serves as the wellness work for the hereditary calculation. The internal advancement plays out the neighborhood seek in a channel way, in which an enhanced estimation of the contingent common data goes about as an autonomous measure for highlight positioning assessing not just the importance of the hopeful element to the yield classes additionally the excess to the effectively chose highlights. The internal and external enhancements participate with each other and accomplish the high worldwide prescient precision and in addition the high nearby inquiry productivity. Trial comes about exhibit both miserly element determination and fantastic grouping precision of the technique on a scope of benchmark information sets.

Bir Bhanu, et al. [11] A hereditary calculation (GA) approach is displayed to choose an arrangement of elements to separate the objectives from the characteristic jumble false cautions in SAR pictures. Four phases of a programmed target recognition framework are created: the harsh target location, include extraction from the potential target locales, GA based component determination and the last Bayesian order. Trial comes about demonstrate that the GA chose a decent subset of components that gave comparative execution to utilizing every one of the elements

H. Chouaib, et al. [12] The system introduces a quick strategy utilizing straightforward hereditary calculations

(GAs) for elements determination. Not at all like customary methodologies utilizing GAs, we have utilized the mix of Adaboost classifiers to assess a person of the populace. In this way, the fitness work we have utilized is defined by the mistake rate of this blend. This approach has been actualized and tried on the MNIST database and the outcomes confirm the viability and the vigor of the proposed approach.

Mohd Saberi Mohamad, et al. [13] We propose a proficient element choice technique that finding and selecting enlightening components from little or high measurement information which most extreme the order exactness. In this work, we apply hereditary calculation to seek out and distinguish the potential educational elements blends for arrangement and afterward utilize the characterization precision from the bolster vector machine classifier to decide the wellness in hereditary calculation.

Chih-Fong Tsai, et al. [14]. The point of this review is to perform include choice and example choice in view of hereditary calculations utilizing distinctive needs to look at the classification exhibitions over various space datasets. The test comes about acquired from four little and vast scale datasets containing different quantities of elements and information tests demonstrate that performing both component and occurrence choice for the most part make the classifiers (i.e., bolster vector machines and k-closest neighbor) perform marginally poorer than highlight determination or case choice separately. Be that as it may, while there is not a significant contrast in classification exactness between these distinctive information pre-handling techniques, the blend of highlight and occurrence determination to a great extent decreases the computational exertion of preparing the classifiers, rather than performing highlight and case choice separately. Considering both classification viability and efficiency, we show that performing highlight choice first and example determination second is the ideal answer for information pre-handling in datamining. Both SVM and k-NN classifiers give comparative classification exactness to the baselines (i.e., those without information pre-preparing). The choices with respect to which information pre-handling undertaking to perform for various dataset scales are likewise talked about.

Tansel Dokeroglu, et al. [15] We propose an arrangement of hearty and versatile cross breed parallel calculations that exploit parallel calculation procedures, developmental gathering hereditary metaheuristics, and canister situated heuristics to get answers for extensive scale one-dimensional BPP cases. An aggregate number of 1318 benchmark issues are analyzed with the proposed calculations and it is demonstrated that ideal answers for 88.5% of these occasions can be acquired with viable streamlining times while taking care of whatever is left of the issues without any than one additional container. At the point when the outcomes are contrasted and the current best in class heuristics, the created parallel half breed gathering hereditary calculations can be considered as one of the best one-dimensional BPP calculations as far as calculation time and arrangement quality.

Zheng Zhao, et al. [16] We propose another "Comparability Preserving Feature Selection" system in an unequivocal and thorough way. We appear, through hypothetical examination, that the proposed structure not just includes many broadly utilized element choice criteria, additionally actually beats their regular shortcoming in taking care of highlight excess. In building up this new system, we start with a traditional combinatorial streamlining detailing for closeness saving element choice, then augment it with an inadequate numerous yield relapse definition to enhance its efficiency and viability. An arrangement of three calculations are conceived to efficiently understand the proposed details, each of which has its own particular points of interest as far as computational multifaceted nature and choice execution. As showed by our broad test think about, the proposed structure accomplishes unrivaled component determination execution and alluring properties.

Ce Zhang, et al. [17] We build up a general hypothesis giving sufficient conditions under which genuine components are ensured to be effectively identified. Unrivaled execution of our technique is shown on a testing connection extraction errand from a huge information set that have both repetitive components and test estimate in the request of millions. We give far reaching correlations best in class highlight choice techniques on a scope of information sets,for which our strategy shows aggressive execution as far as running time and exactness. In addition, it additionally yields generous speedup when utilized as a pre-preparing venture for most other existing techniques.

Hanchuan Peng, et al. [18] We introduce a two-arrange include choice calculation by consolidating mRMR and other more advanced component selectors (e.g., wrappers). This permits us to choose a reduced arrangement of prevalent elements with ease. We perform broad exploratory examination of our calculation and different strategies utilizing three unique classifiers (guileless Bayes, bolster vector machine, and straight separate investigation) and four distinct information sets (written by hand digits, arrhythmia, NCI growth cell lines, and lymphoma tissues). The outcomes affirm that mRMR prompts to promising change on highlight choice and arrangement precision.

T. Hilda, et al. [19] Include choice is a powerful procedure for dimensionality lessening and a fundamental stride in fruitful information mining applications. It is a procedure of selecting a subset of components from the applicant set of elements as indicated by specific criteria. The principle objective of managed learning is discovering highlight subset that produces higher characterization precision. The proposed strategy is to choose an ideal arrangement of components by utilizing Genetic Algorithm that has been done in parallel by utilizing MapReduce system. The resultant components will be offered it to the K-Nearest Neighbor classifier. The wellness of exactness will be assessed utilizing K-NN. Results are approved utilizing the Datasets taken from the UCI machine learning vault. The outcomes show that the Parallel GA creates high exactness than different techniques.

3. Proposed Work

The main idea behind this chapter is to give brief idea about performing feature selection for the group of features. The overall Design approach is basically divided into several steps. The first step is input data sets is used which is available from UCI machine learning repository datasets for feature selection. The three datasets are used i.e Ionosphere, Wdbc, Statlog (heart).The data sets which is being used have not provide any group information. Creating the group of features is the second step. The group of features is created by dividing the feature randomly. The size of group is depending on the user choice. This step gives the group of feature.

Next step is performing feature selection on group of features, we focus on the problem where feature possessing some group structure, the step gives the optimal subsets of features. The validation is needed on the selected feature in order to evaluate whether the features are optimal or not classification is required. The KNN classifier is applied to evaluate the performance of selected.

Steps involved in genetic feature selection

- The random generation of population is done by fitness level Evaluation
- Depending upon the fitness the two parents are selected
- The new offspring"s (children) are generated after crossover of parents.

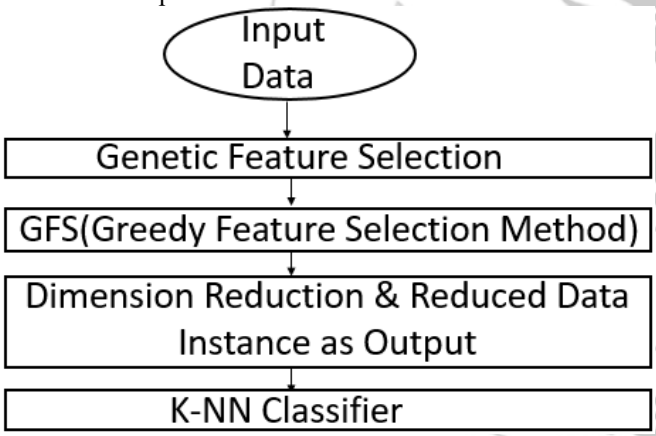

Figure: Architecture of proposed model

- Mutation is done at every single point.
- The placement of new children in population
- The algorithm gets running by new population.
- The conditions are satisfied then the algorithm stop otherwise it returns the best solution

4. Conclusion and Future Work

GA has been found a decent strategy for improvement and looking for quality arrangement. The element choice is a stage to choose an ideal component from unique list of capabilities. It is a productive technique to lessen dimensionality and evacuate undesirable information. Gather structure is an accumulation of elements. This serves to builds exactness and reductions computational time. This examination extends presented another technique for highlight having bunch structure called hereditary element determination (GFS). This expands the order precision and demonstrates the adequacy of our technique. Additionally, it gives great bring about selecting ideal component subset from a gathering of elements. Future work is to focus on Hybrid element choice techniques like Cuckoo Search with GA can likewise be demonstrated and tested. Likewise, the utilization of Hadoop MapReduce will lessen the computational time of highlight determination demonstrate.

References

[1] J. Wang, Z.Q. Zhao, X. Hu, Y.M. Cheung, M. Wang, and X. Wu, "Online group feature selection," in IJCAI, 2013, pp. 1757–1763.
[2] Haiguang Li, Xindong Wu, Zhao Li, Wei ding "Group feature selection with streaming features," 2013 IEEE 13th international conference on data mining.
[3] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society, vol. 68, no. 1, pp. 49–67, 2006.
[4] H.Yang, Z. Xu, I. King, and M. R. Lyu, "Online learning for group lasso," in ICML, 2010, pp. 1191–1198.
[5] Lei Yuan, Jun Liu, and Jieping Ye, "Efficient method for overlapping group lasso," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 9, September 2013.
[6] S. Xiang, X. T. Shen, and J. P. Ye, "Efficient sparse group features election via nonconvex optimization," in ICML, 2012.
[7] M. Ramaswami and R. Bhaskaran (2009), "A Study on Feature Selection Techniques in Educational Data Mining", Journal of Computing.
[8] Sunita Beniwal, Jitender Arora (2012), "Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181.
[9] Huan Liu, Hiroshi Motoda, Rudy Setiono and Zheng Zhao, "Feature Selection: An Ever-Evolving Frontier in Data Mining", In the Fourth Workshop on Feature Selection in Data Mining, JMLR: Workshop and Conference Proceedings 10: 413.
[10] Jinjie Huang, Yunze Cai, Xiaoming Xu (2007), "A hybrid genetic algorithm for feature selection wrapper based on mutual information", Pattern Recognition Letters 28, pp.1825– 1844.
[11] Bir Bhanu, Yingqiang Lin (2003), "Genetic algorithm based feature selection for target detection in SAR images", Image and Vision Computing 21, pp. 591–608, 2003.
[12] H. Chouaib, O. Ramos Terrades, S. Tabbone, F. Cloppet, N. Vincent (2005), "Feature selection combining genetic algorithm and Adaboost classifiers", IEEE transactions.
[13] Mohd Saberi Mohamad, Safaai Deris, Safie Mat Yatim, Muhammad Razib Othman (2004), "Feature selection method using genetic algorithm for the classification of small and high dimension data", First International Symposium on Information and Communications Technologies. October 7-8.

[14] Chih-Fong Tsai, William Eberle, Chi-Yuan Chu (2013), "Genetic algorithms in feature and instance selection", Knowledge-Based Systems 39, pp.240–247.
[15] Tansel Dokeroglu, Ahmet Cosar (2014), "Optimization of one-dimensional bin packing problem with island parallel grouping genetic algorithms", Computers & Industrial Engineering.
[16] Zheng Zhao, Lei Wang, S, Huan Liu, and Jieping Ye,On "similarity preserving feature selection," IEEE transactions on knowledge and data engineering, vol. 25, no. 3, march 2013.
[17] Ce Zhang, Hung Ngo, Xuan Long Nguyen "Parallel Feature Selection inspired by Group Testing," IEEE transactions on knowledge and data engineering, vol. 36, no. 4, march 2013.
[18] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information criteria of Max dependency, max relevance and min redundancy," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 8, august 2012.
[19] T. Hilda and Dr. R.R.Rajalaxmi "Effective Feature Selection for Supervised Learning Using Genetic Algorithm". IEEE SPONSORED 2'ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS (ICECS ,2015).

Author Profile

Rizwan Ahmed khan is a M.Tech student from G.H Raisoni College of Engineering, Nagpur (An Autonomous Institute Affiliated to RTM Nagpur) he has completed his B.E. from Dr. Babasaheb Ambedkar Marathwada University Aurangabad in the year 2012. He is intrested in data mining, and has greater interest in Networking.

Indu Mandwi is an Assistant Professor at the G.H Raisoni College of Engineering, Nagpur (An Autonomous Institute Affiliated to RTM Nagpur) She has completed a Post graduate degree in Embedded System and Computing at G. H. Raisoni College of Engineering, Nagpur between 2010 and 2012. She has done B.E. In Computer Science and Engineering from Samrat Ashok Technological Institute, Vidisha in 2006. After completing B.E. She took up her first post as an Assistant Professor at the Radharaman Institute of Technology and Science, Bhopal. She has 10 years of working experience. Her area of interest includes Database management system, Embedded System, Operating system, Data mining.