IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Multi-Objective Optimization for Feature Selection using Pareto Front

Bhuvaneswari Ragothaman[1], Dr. B.Sarojini[2]

Research Scholar, Department of Computer Science, Avinashilingam University, Coimbatore, India[1]

Assistant Professor, Department of Computer Science, Avinashilingam University, Coimbatore, India[2]

Abstract: The huge amount of data is collected from the real world and stored in the databases. The extraction of useful information from the large data is a tedious task. The effective extraction of the data can obtain useful information. The classification of the large data regarding their features is difficult. The feature selection for the particular data set in order to classify the data should be performed in clear way to obtain better classification. Different algorithms are used in the data mining to obtain the classification in a better manner with effective feature selection process. The problem confronted by the classification process is due to the presence of redundant data and irrelevant data. In feature selection process the data present in the set may not support with similar feature as some points having irrelevant feature then specified. The existing classification process can't compete with the minimum local deviation and they include high computation cost. In order to achieve better classification of the data to obtain the appropriate result this work includes implementation of the Pareto Optimization along with optimization algorithm to extract the feature subset. The Pareto front is used to determine non-dominated feature subset in order to compute the feature extraction in minimal number and to increase the classification accuracy.

Keywords: Classification, Feature Selection, Pareto Optimization, Optimization Algorithm.

I. INTRODUCTION

In the healthcare field, data are collected from various areas like patient personal data, diagnostic data, lab data, observation data, machine generated data and etc. Those data are large in amount with redundancy and missing values. Here feature selection is the main task which is to be done for the selection of relevant features without any duplication or noise data [5]. The selected feature subset increases the classification accuracy of the dataset. Feature selection for the dataset can be done by two approaches like filter and wrapper. Filter Approach selects the features by using some threshold value to form a feature subset. Wrapper Approach frames a new feature subset with the original features. The feature subset selected will help in maximizing the classification accuracy.

There exist various optimization algorithms which help in selecting the feature subset. Pareto Optimization can be implemented along with the feature selection algorithm to select the feature subset and to minimize the size of the feature subset. This feature subset selected can be used for the classification of the dataset and can improve the accuracy of classification [4].

II. RELATED WORKS

Xue and et al. [1] have proposed Multi Objective Approach for feature selection using PSO Algorithm, in which feature subset determines a non dominated feature subset Non-dominated Sorting PSO Feature Selection Algorithm and dominated feature subset using Crowding Mutation Dominated PSO Feature Selection Algorithm. In first algorithm, the Pareto front value is calculated and with that non dominated solution is calculated using PSO algorithm. In CMDPSOFS, with the non dominated solution along with crowding and mutation the feature subset is being determined. Here wrapper approaches like LFS and GSBS are used for feature selection. With the selected feature subset the classification of the data is done using KNN classifier.

Emrah and et al. [2] have proposed Multi Objective Artificial Bee Colony Algorithm for feature selection using mutual information and Artificial Bee Colony Algorithm. Using the fuzzy mutual information between the features along with the working and non working bees the non dominated feature sub set is being determined.

Kalyanmoy and et al [3] precision amplification and many-sided quality minimization are the two fundamental objectives of a fluffy master framework based microarray information arrangement. In past Genetic Swarm Algorithm (GSA) approach has enhanced the order exactness of the fluffy master framework at the expense of their interpretability. The if-then guidelines created by the GSA are long and complex which is troublesome for the doctor to get it. To address this interpretability-exactness tradeoff, the principle set is spoken to utilizing whole number numbers and the errand of guideline era is dealt with as a combinatorial advancement undertaking. Insect colony optimization (ACO) with neighborhood and worldwide pheromone updating are connected to discover the fluffy allotment taking into account the quality expression values for creating more straightforward

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

guideline set. Keeping in mind the end goal to address the undefined and persistent expression estimations of a quality, this paper utilizes fake honey bee state (ABC) calculation to advance the purposes of participation capacity. Shared Information is utilized for identification of useful qualities. The execution of the proposed cross breed Ant Bee Algorithm (ABA) is assessed utilizing six quality expression information sets

III. PROPOSED SYSTEM

Feature Selection is a process of selecting the relevant features from the dataset to attain the optimal solution for solving the problem. Using this reduced feature subset the classification accuracy can be increased but still there exists some error rate in the classification. This error rate can be reduced using some of the optimization algorithms like Genetic Algorithm, Particle Swarm Optimization Algorithm, Ant Colony Optimization Algorithm or Artificial Bee Colony Algorithm. When there is more than one objective to be solved then it is said to be the multi objective optimization.

In the case of Feature Selection problem, the two main objectives to be solved are,

i. To minimize the number of features in the feature subset
ii. To maximize the accuracy of the feature's classification.

To attain both the objectives of Dimensionality Reduction Pareto Front can be used.

A. Pareto Optimization

Pareto Optimization is a mathematical process of solving the multi objective problems in step by step. Pareto Front is the feasible value or the sustain value which gives the correct feature subset with maximum classification accuracy. To calculate the Pareto Front value Pareto Curve is to be generated with the non dominated features present in the dataset.

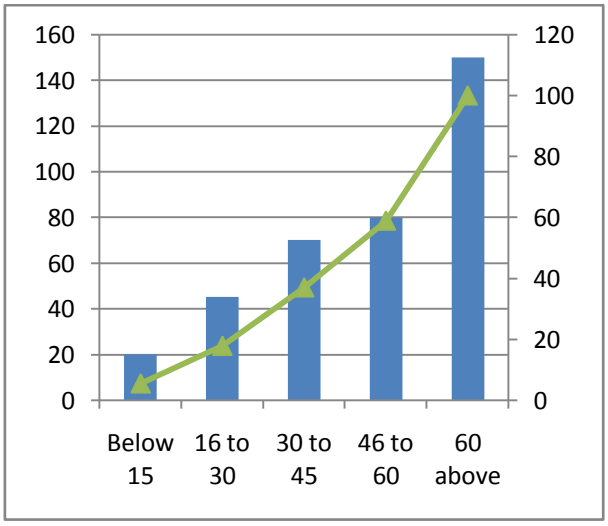

Fig 1 Pareto Curve

To draw the Pareto Curve, first the feature's weight are being calculated using the formula,

$F = [f_1(x), f2(x), f3(x) ... fm(x)]$

$G_i(x) <= 0; I = 1, 2 \ldots m$

$H_i(x) <= 0; I = 1, 2 \ldots m$

where, x is decision variable, fi(x) is function of x, $G_i(x)$ is inequality constrain, $H_i(x)$ is equality constrain [1]. The Pareto curve is shown in fig1.

With these mathematical formulas, the features are differentiated into dominated and non dominated features. Pareto curve is being drawn only using the non dominated features of the dataset [6]. Then the Pareto Front value is being calculated from the values of the Pareto Curve. The Pareto front value will be the average value of the Pareto curve.

B. Determining Feature Subset

Feature selection is connected to inputs, unsurprising properties, or to states in a section. At the point when scoring for highlight determination is finished, just the traits and states that the calculation chooses are incorporated into the model-building process and can be utilized for expectation. On the off chance that you pick an anticipated characteristic that does not meet the edge for highlight determination the quality can at present be utilized for forecast, yet the expectations will be construct exclusively in light of the worldwide measurements that exist in the model.

Feature selection for the most part influences the preparation period of classification. In the wake of creating components, rather than preparing information with the entire components to the learning calculation specifically, highlight determination for arrangement will first perform highlight choice to choose a subset of components and afterward prepare the information with the chose elements to the learning calculation. The element determination stage may be autonomous of the learning calculation, similar to channel models, or it may iteratively use the execution of the learning calculations to assess the nature of the chose highlights, similar to wrapper models. With the at last chose highlights, a classifier is instigated for the expectation stage. Typically highlight choice for grouping endeavors to choose the insignificantly measured subset of components as indicated by the accompanying criteria,

- The characterization accuracy does not fundamentally diminish; and
- The subsequent class conveyance, given just the qualities for the chose elements, is as close as could be expected under the circumstances to the first class appropriation, given all features.

Using any optimization algorithms like Genetic Algorithm, Ant Colony Optimization Algorithm, Particle Swam Optimization Algorithm, Artificial Bee Colony Algorithm an optimistic feature subset can be determined. The feature subset selected may cause error in classification [7]. By comparing the Pareto front value to

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

the features selected the Non-Dominated Feature Subset is selected with minimal number of features.

C. Classification Algorithm

There are various classification algorithms available for classification of the dataset like, Random Forest, KNN, SVM and etc. The performance of the Non-Dominated Feature Subset is checked by the parameters of the classification algorithm like, precision, recall, accuracy, f-measure and error rate of the classification problem. Those are calculated using the confusion matrix values shown in table 1.

Table1 Confusion Matrix

Actual Values	Predicted	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

IV. CONCLUSION

The high dimensional data classification is obtained in the accurate manner in this work by using the feature selection process. The data set used contains the irrelevant information where they are get removed by the preprocessing stage. The objective of the Pareto optimization is to select the feature that helps in better extraction of relevant data. This value is also concentrated in the non-dominant solutions in the multi objective feature selection. The reduced dimensionality with the original features helps in the feature selection process. The selected feature is classified for the correct label observation by using the classification Algorithm. The better classification of the data is obtained with less number of features set and less computational time is observed for the classification due to the selection of relevant feature.

REFERENCES

1. Xue, Member, IEEE, Mengjie Zhang, Senior Member, IEEE, and Will N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 6, DECEMBER 2013.
2. Emrah Hancer, Bing Xuey, Mengjie Zhangy, Dervis Karaboga and Bahriye Akay, "A Multi-Objective Artificial Bee Colony Approach to Feature Selection Using Fuzzy Mutual Information", IEEE,2015.
3. Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan," A Fast and Elitist Multi objective Genetic Algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation, VOL. 6, NO. 2, April 2002.
4. Chia-Hung Hsu and Chia-Feng Juang, "Multi-objective Continuous-Ant-Colony-Optimized FC for Robot Wall-Following Control" IEEE Computational intelligence magazine August 2013.
5. Luiz S. Oliveira, Marisa Morita, and Robert Sabourin, "Feature Selection for Ensembles Using the Multi-Objective Optimization Approach", Springer-Verlag Berlin Heidelberg 2006.
6. Ashish Ahuja, Sanjoy Das, and Anil Pahwa," An AIS-ACO Hybrid Approach for Multi-Objective Distribution System Reconfiguration", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 3, AUGUST 2007.
7. Wenping Zou, Yunlong Zhu, Hanning Chen and Beiwei Zhang,"Solving Multi objective Optimization Problems Using Artificial Bee Colony Algorithm", Hindawi Publishing Corporation Discrete Dynamics in Nature and Society Volume 2011, Article ID 569784.