International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A survey on Genetic Algorithm based Gene Subset Selection for Disease Prediction: Underlying Concepts & Approaches

Deekshita S[1], Divya Balaji[2], S Harshita[3], J Karthiyayini[4]
[1, 2, 3, 4,] *Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore, India*

Abstract*: Disease prediction is of critical importance with respect to advancements in health-care as well as technology. With the advent of big data and analytics, it has become feasible to develop methods to diagnose and predict diseases in cost effective ways. Concepts such as Whole Genome Sequencing has revolutionized healthcare and the way data can be used to tackle diseases. Evolutionary algorithm(EA) or Genetic Algorithm is one such method; it is a subset of evolutionary computation which is a generic population based optimization method. EA is inspired by biological methods such as reproduction, mutation etc. It differs from other optimization methods in its randomness, crossover, mutation and selection. Micro-array data is used extensively for feature subset selection of genes. However, micro-arrays have a large number of features and comparatively small number of samples making disease prediction challenging. Thus, it is important to identify and use the best gene subset selection technique. A comparative approach to the existing methods and its latest versions is studied to analyze the methods that produces the superlative result. This is done with the help of experimental results obtained and its accuracy with respect to disease prediction.

Index Terms: Genetic algorithm, gene subset selection, disease prediction, optimization methods.

I. INTRODUCTION

A. Big Data

According to Healthcare IT news, in the United States alone, 400,000 people die from preventable medical errors each year. However, this alarming number can be brought down with the use of one the most debated topic of all: big data. By analyzing large healthcare datasets, analysis and insights can be drawn with the help of big data that will help prevent medical errors and solve healthcare's biggest problems. In addition, the association of big data with artificial intelligence can be used to scan cloud-based data to help prevent misdiagnosis and transcription errors. From a higher view, big data seems just another analysis tool, however its depth and efficiency cuts down on both the time & resources to solve the biggest problems in the healthcare industry. More so, hospitals, clinics and healthcare providers can provide a safer and more efficient patient experience, improve outcomes & enable better communication. According to Hermon [1], big data analytics can bring the revolution in healthcare industry. This data in healthcare provide opportunity to perform predictive analysis.

B. Genetic Algorithm

Genetic algorithm(GA) is a stochastic search method, however, the genetic algorithm is inspired not by physics but by biology - specifically the evolutionary processes such as reproduction, selection, mutation and crossover. The GA approach has the following features- an abstraction representation of the system, a slightly different version of the system is developed and the whole population is improved rather than a single entity. A function produces a single value that characterizes the 'fitness' of the fitness function. However there involve limitations to genetic algorithms. Obtaining solutions to high dimensional, multimodal problems requires expensive evaluations of the fitness function. GAs do not scale well with respect to complexity. Thus it is important to ensure that the other processes involved use the best technique, such as the gene subset selection methods for effective disease prediction.

C. Feature Selection

Feature selection is aimed at identifying a minimal-sized subset of features that are relevant to the target concept [2]. The objective of feature selection is threefold: improving the prediction accuracy, providing faster and more cost-effective prediction, and providing a better understanding of the underlying process that generated the data [3]. A feature selection method generates different candidates from the feature space and assesses them based on some evaluation criterion to find the best feature subset [2].

The researches has proofed feature selection is very effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results[4],[5].In recent years many techniques have proposed in the literature for gene feature selection and cancer classification.

D. Micro-Arrays

Microarray technology is used in data mining algorithms for finding cancerous diseases. Xiaosheng Wang and Osamu gotoh[6] demonstrate that one major problem in pertaining the gene expression profiles to cancer classification and prediction is that the number of gene features greatly surpass the number of samples. Some researchers have shown that a small collection of genes selected correctly can lead to good classification results. Therefore gene feature selection is very essential in cancer classification. Various methods of selecting informative gene groups to conduct cancer classification have been proposed. Many methods are used for classification such as wrapper, filter and embedded methods.

E. SPEA & SPEA2

The Strength Pareto Evolutionary Algorithm (SPEA) is used to locate and maintain a front of non-dominated solutions, ideally a set of Pareto optimal solutions. This is achieved by using an evolutionary process (with surrogate procedures for genetic recombination and mutation) to explore the search space, and a selection process that uses a combination of the degree to which a candidate solution is dominated (strength) and an estimation of density of the Pareto front as an assigned fitness. An archive of the non-dominated set is maintained separate from the population of candidate solutions used in the evolutionary process, providing a form of elitism.[7] All existing traditional data mining techniques have not capability to perform better classification for big data in health-care domain.

Therefore, there is need to merge the distinct techniques together to perform better classification [8]. One popular approach for requiring relationship between data is Association. However, there exist shortcomings in SPEA, which are solved by the enhancements in SPEA2. SPEA2 provides a better distribution of points, especially when the number of objectives increases.

Furthermore, it became obvious that it is necessary to trace the performance over time to keep track of the dynamic behavior of the algorithms. Specifically, algorithms are likely to differ in convergence velocity or reveal effects such as premature convergence or stagnation, which cannot be seen from a static analysis after some arbitrary running time.[9]

II. UNDERLYING CONCEPTS & APPROACHES

A. Gene subset Selection

Extracting relevant information from micro-array data is a very complex task due to the characteristics of the data sets, as they comprise a large number of features while few samples are generally available. In this sense, feature selection is a very important aspect of the analysis helping in the tasks of identifying relevant genes and also for maximizing predictive information. Better cancer outcome prediction results were obtained using the GA framework noting that this approach, in comparison to the SFS one, leads to a larger selection set, uses a large number of comparison between genetic profiles and thus it is computationally more intensive. Also the GA framework permitted to obtain a set of genes that can be considered to be more biologically relevant. This study shows that if prediction accuracy is the objective, the GA-based approach lead to better results respect to the SFS approach, independently of the classifier used. Regarding classifiers, even if C-MANTEC did not achieve the best overall results, the performance was competitive with a very robust behaviour in terms of the parameters of the algorithm

B. Clustering Algorithms

Clustering algorithms present a number of problems that make their application difficult and/or restrict the amount of knowledge that could be extracted from the analysis. One of the problems of the usual application of cluster analysis for the exploration of a data set is its focuses on the discovering of single partitions that best fit the data. In contrast, in several practical cases the data can present distinct underlying structures, each one characterized by a different partition. A selection strategy, based on the corrected Rand index is presented that aims at recommending, as final solutions for Pareto-based multi-objective genetic algorithm approaches, a subset of partitions from the Pareto front.

In order to test our strategy, we develop a study of case in which we apply the strategy to the sets of solutions obtained with the Multi-Objective Clustering Ensemble algorithm (MOCLE) in the context of several data sets. One of the advantages of Pareto-based multi-objective genetic algorithms for clustering, when compared to classical clustering algorithms, is that, instead of a single solution (partition), they give as an output a set of solutions (approximation of the Pareto front or Pareto front, for short).

C. CAST (Clustering Affinity Search Technique)

is a famous clustering algorithm, which is widely used in clustering the biological data. Two algorithms, namely Calculation-On-Demand CAST, abbreviated as COD-CAST and Calculation-On-Demand CAST with GPU, abbreviated as COD-CAST-GPU, respectively is proposed. The first proposed COD-CAST algorithm is a refined CAST algorithm that can process large amount of objects more efficiently in terms of execution time. The proposed COD-CAST-GPU algorithm can utilize the GPU and the individual memory of graphics card to accelerate the COD-CAST. The experimental results show that the proposed algorithms deliver excellent performance in terms of execution time and required memory. The advances in nanometer technology and integrated circuit technology enable the graphics card to attach individual memory and one or more processing units, named GPU, in which most of the graphing instructions can be processed parallelly. The computation resource can be used to improve the execution efficiency of not only graphing applications but other time consuming applications like data mining.

A new evolutionary method for the cluster validation index (CVI), namely eCVI. The method learns CVI from the generated training data set using the genetic programming (GP), and then outputs the optimal number of clusters after taking parameters of a test data set into the learned CVI. Each chromosome encodes a possible CVI as a function of the number of clusters, density measure of clusters, and some random factors. Fitness function evaluating each candidate is defined by the difference between the actual number of clusters from training data set and the number of clusters computed by the current CVI. Fitness function evaluating each candidate is defined by the difference between the actual number of clusters from training data set and the number of clusters computed by the current CVI. The proposed eCVI is reliable and robust in various types of data sets because of the adaptive nature of GP. Experimental results provide grounds for the dominance of eCVI over several widely-known CVIs.

D. Genetic Algorithm

A new fitness assignment scheme to evaluate the Pareto-optimal solutions for multi-objective evolutionary algorithms proposed that Domination Power of an individual Genetic Algorithm (DOPGA) method can order the individuals in which each individual solution has a unique rank. A multi-objective problem can be treated as if it were a single-objective problem without drastically deviating from the Pareto definition. In DOPGA, relative position of a solution is embedded into the fitness assignment procedures. The performance comparison of the algorithm is with two benchmark evolutionary algorithms (Strength Pareto Evolutionary Algorithm (SPEA) and Strength Pareto Evolutionary Algorithm 2 (SPEA2)) on 12 unconstrained bi-objective and one tri-objective test problems. DOPGA significantly outperforms SPEA on all test problems.

DOPGA performs better than SPEA2 in terms of convergence metric on all test problems. Also, Pareto-optimal solutions found by DOPGA spread better than SPEA2 on eight of 13 test problems. A Set of methods that uses a genetic algorithm for automatic test-data generation. These methods will take the test populations as an input and then evaluate the test cases. Software testing using Genetic Algorithms become efficient even with increasing number of test cases that increases the efficiency and process time of Software testing. Hence, random testing methods becomes inefficient as data points do not have a dependence with time and code becomes complex.

When using Modelling techniques- Naive Bayes Rule and Genetic algorithm, that uses the concepts of Conditional Probability and Genetic Search method respectively, which predict the risk level of heart diseases plan to develop a Intelligent Heart Disease Decision Support System based on Genetic Algorithms and Fuzzy Logic. The use of Bayes' Rule is quite effective but improvements can be made in terms of the number of attributes used. Which concludes that Genetic algorithms are more effective.

The study has implemented K-Means and GA for dimensionality reduction and SVM to classify the diabetes data-set. Which Implements K-Means for removing the noisy data and Genetic Algorithms for finding the optimal set of features by using Support Vector Machine(SVM) as a classifier. It's observed that the SVM classification accuracy of the proposed method is better than the Modified K-Means and SVM reported.

E. SPEA & SPEA2

Evolutionary algorithms (EAs) are often well-suited for optimization problems involving several, often conflicting objectives. Four multi objective EAs are compared quantitatively where an extended 0/1 knapsack problem is taken as a basis. A new evolutionary approach to multi-criteria optimization, the strength Pareto EA (SPEA), that combines several features of previous multi objective EAs in a unique manner. It is characterized by (a) storing non dominated solutions externally in a second, continuously updated population, (b) evaluating an individual's fitness dependent on the number of external non dominated points that dominate it, (c) preserving population diversity using the Pareto dominance relationship, and (d) incorporating a clustering procedure in order to reduce the non-dominated set without destroying its characteristics.

The proof-of-principle results obtained on two artificial problems as well as a larger problem, the synthesis of a digital hardware-software multiprocessor system, suggest that SPEA can be very effective in sampling from along the entire Pareto-optimal front and distributing the generated solutions over the tradeoff surface. SPEA clearly outperforms the other four multi-objective EAs on the 0/1 knapsack problem. A new evolutionary approach to multi criteria optimization, the strength Pareto EA (SPEA), that combines several features of previous multi-objective EAs in a unique manner. SPEA clearly outperforms the other four multi-objective EAs on the 0/1 knapsack problem. SPEA2 is more reliable for BPP problem as against the other methods used. As it is used for Burnable Poison Placement (BPP) using SPEA2 of a nuclear reactor core. Nodal expression code is used along with SPEA2. This is an enhanced approach with better proved results hence, convenient and accurate results are obtained. Also it is an comparative results- basic approach of SPEA2 method.

Applying SPEA2 and comparing the results obtained against a decision support tool, RAILER to evaluate whether an additional cost solution is worth the increase rail condition. That has employed a well-benchmarked Pareto-based MOEA (SPEA2) to the Fort Smith rail repair problem. The solutions generated dominate the RAILER solution. Utilizing the RAILER prioritized worst first scheme, the best condition achieved is a score of 768 at a cost of $12.9M.But lower cost using SPEA2. SPEA2 employs an elitist strategy in that it maintains an external archive of solutions and exclusively produces new solutions from the archive members. SPEA2 also utilizes a nearest neighbor density estimation technique that is used to differentiate solutions and maintain solution spread (diversity) within the archive. It also provides Increased flexibility.

An improved SPEA2 algorithm with adaptive selection of evolutionary operators. Multi objective evolutionary operators including the simulated binary crossover, polynomial mutation, and differential evolution operator are employed to enhance the convergence performance and diversity of the SPEA2.Simulation results on the standard benchmarks show that the proposed algorithm outperforms SPEA2,NSGA-II and PESA-II. The need to improve the performance of AOSPEA by making use of the adaptive scheme to mutation operator and verifying its efficiency through a comparison with other types of MOEAs. More than two objectives in the MOPs are best studied. AOSPEA utilized to solve the multi objective job shop and flow shop scheduling problems. The selective way makes the proposed algorithm achieves the optimal values faster. A minimum selection probability is also set to avoid some operators which would have strong search ability in the remaining process of the algorithm. The experimental results verify these points. The strength of the AOSPEA is not obvious while optimizing the instances with high dimensions. Therefore, no reliable method to set the value of minimum selection probability is acquired.

Comparison of SPEA which is a relatively recent technique for finding the Pareto-optimal sets, and SPEA2 which is an improved version of SPEA, and other two modern elitist methods methods, PESA and NSGA-II, on different test problems yields promising results. As SPEA2 performs better than SPEA in all test problems. PESA has faster convergence due to its higher elitism intensity, but has difficulties on some problems due to the inability to keep boundary solutions. SPEA2 is better than PESA and NSGA-II. Comparative study shows many differences only for more objectives than two. Challenge for the archiving strategies and for the algorithms to maintain a good distribution of solutions, in higher dimensions.

III. PROPOSED WORK

Using SPEA-II and Cluster Affinity Search Technique(CAST) for the gene subset selection, the project intends to identify the gene that helps predict the disease based on the information obtained by applying the two genetic algorithms. The project adopts a backwards approach wherein the shortcomings are used as a foundation to advance in the direction of the final result, which is to obtain gene subsets that help predict diseases. CAST is a fast and practical algorithm that uses a statistical approach to the gene cluster involved in terms of its closeness or distance to a particular cluster under consideration. The above mentioned algorithms and tools are to be implemented in Python to obtain the objective of the project.

IV.CONCLUSION

Due to rapid enhancement in big data prediction and analysis, healthcare domain has got a valuable attention from recent few years. Existing traditional machine learning approaches take much time in computation when data set volume increase. The Adjusted Rand index works well in SPEA2 too as the previous RAND index method adopts different ways to do selection and population maintenance in solving optimization problems. Clustering quality is objectively evaluated using the Adjusted Rand Index, an external measure of clustering quality that is a generalization of the Rand Index. The Rand indices are based on counting the number of pair wise co-assignments of data item. In this case, instead of using Adjusted RAND Index, we use CAST algorithm. The algorithm works by both adding and removing genes from a cluster, each time adjusting the affinities of the genes to the current cluster, and continuing

this process until no further changes can be made to the current cluster. By observing the conclusions drawn by different clustering algorithms, CAST provides a better way of creating and removing clusters for disease prediction.

REFERENCES

[1] R. Hermon and P. A. Williams, “Big data in healthcare: What is it used for?” 2014.

[2] M. Dash, H. Liu, "Feature selection for classification", Intell. Data Anal., vol. 1, no. 3, pp. 131-156, 1997.

[3] Guyon, A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[4] H. Almuallim and T. G. Dietterich. “Learning boolean concepts in the presence of many irrelevant features,” Artificial Intelli-gence, vol. 69, no. 1-2, pp. 279–305, 1994.

[5] D. Koller and M. Sahami, “Toward optimal feature selection,” In Proceedings of the Thirteenth International Conference on Ma-chine Learning, pp. 284–292, 1996

[6] Xiaosheng Wang1 and Osamu gotoh” A Robust Gene selection Method for Microarray-based cancer Classification” Cancer Informatics 2010:9 15–30.

[7] Zitzler and M. Laumanns and S. Bleuler, "A Tutorial on Evolutionary Multiobjective Optimization", in Metaheuristics for Multiobjective Optimisation, pages 3–37, Springer, 2004.

[8] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen,“Application of classification techniques on development an early-warning system for chronic illnesses,”Expert Systems with Applications,vol. 39, no. 10, pp. 8852–8858, 2012

[9] SPEA2: Improving the Strength Pareto Evolutionary Algorithm Eckart Zitzler, Marco Laumanns, and Lothar Thiele Computer Engineering and Networks Laboratory (TIK) Department of Electrical Engineering Swiss Federal Institute of Technology (ETH) Zurich ETH Zentrum, TIK-Report 103 May 2001(Errata added Sept 27, 2001)