World Applied Sciences Journal 31 (6): 1109-1117, 2014
ISSN 1818-4952
© IDOSI Publications, 2014
DOI: 10.5829/idosi.wasj.2014.31.06.455

Applications of ANNs, SVM, MDR and FR Methods in Bioinformatics

[1]*Seyed Nasirodin Mirmohammadi,* [2]*Majid Shishehgar and* [3]*Fereshteh Ghapanchi*

[1]University of Shiraz, Iran
[2]Isfahan University of Technology, Iran
[3]AzadUniversity, Mobarakeh Branch, Iran

Abstract: Bioinformatics, or computational biology, is an interdisciplinary field of study for interpreting biological data using computer science. The importance of this new field of researchwill grow as we continue to generate and integrate large quantities of genomic, proteomic and other data. An interesting area of research in bioinformatics is the application and development of machine learning techniques to solve biological problems. Several efforts are being made by the computer scientists and statisticians to design and implement algorithms and techniques for efficient storage, management, processing and analysis of biological databases. In this research we review the application of some of the most popular and applicable machine learning techniques in the field of bioinformatics. The four presented algorithms, including ANNs (Artificial Neural Networks), SVM (Support Vector Machines), MDR (Multifactor Dimensionality Reduction) and RF (Random Forrest) are used to address some problems faced in computational biology. Therefore, first we introduce the bioinformatics and its potential problems, then the four methods have fully surveyed separately with their application in addressing some specific biological problems and finally a comparison table have presented which shows the pros and cons and their application of the four machine learning techniques.

Key words: Bioinformatics · Random Forest · Data mining · Classification · Gene-geneinteraction

INTRODUCTION

Bioinformatics is an emergingdiscipline that addresses the need to manage and interpret the data that in the past decade was massively generated by genomic research. This discipline represents the convergence of genomics, biotechnology and information technology and encompasses analysis and interpretation of data, modeling of biological phenomena, such as deoxyribonucleic acid (DNA) sequencing which is one of the most important platforms for the study of biological systems today [1] and development of algorithms and statistics. DNA sequence contains genes and gene comprises genic and inter-genic regions. Ribonucleic acid (RNA) translation from DNA is an important and critical task because exact identification of protein helps in knowing information regarding protein structure and cell functions [2].

A particular active area of research in bioinformatics is the application and development of artificial intelligence techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Therefore, we see a great potential to increase the interaction between machine learning and bioinformatics.

Bioinformatics or Computational Biology: Generally in recent years, with development of genetic science, the gene has become a main factor in planning of gene performance. Genes influence all human diseases and yet much of the genetic landscape of many common diseases is still uncharacterized. Genome-wide association studies

Corresponding Author: Seyed Nasirodin Mirmohammadi, Mater of Information Technology Graduate, E-Learning department of University of Shiraz, Iran. Tel: +98-9133707637.

(GWAS) using single nucleotide polymorphisms (SNPs) have been extensively used to uncover genetic architecture [3] by testing variants individually for association with particular diseases or traits [4,5]. However, GWAS have explained only a small proportion of the genetic variation underlying disease [3,6] For common diseases the effect of an individual SNP on disease susceptibility is generally small and emerging evidence suggests that many low-penetrance variants interact multiplicatively [7] with increasing numbers of risk alleles contributing to significantly elevated disease risks [8].

Therefore, it is likely that much of the genetic variation underlying common diseases arises through interactions between many genes and environmental factors; a form of epistasis [9]. Thus the identification of individual disease-related SNPs may be less useful for disease prediction than the identification of the epistatic relationships underlying genetic disease. The term epistasis has been used to refer to at least two phenomena which may be related in complex ways. Biological epistasis, which occurs at the cellular level, corresponds to the physical interactions amongst biomolecules in gene regulatory networks and pathways that impact on phenotype. Therefore, the impact of a gene on an individual's phenotype depends on one or more additional genes. Alternatively, statistical epistasis reflects differences in biological epistasis among a population of individuals: the deviation from additives within a statistical model of the relationship between multiple genotypes and phenotype (s) at a population level [3, 10, 3] Presents conceptual relationships between biological and statistical.

Phillips in 2008 has suggested that epistasis can be split into three categories [11]:

1) Compositional epistasis, 2) functional epistasis and 3) statistical epistasis. Compositional epistasis is introduced to represent the traditional definition of epistasis as the blocking of the effect of an allele by an allele at another locus. However defined, the relationships between biological and statistical forms of epistasis are complex and statistical interaction does not necessarily reflect interaction on a biological level [12].One of the major problems associated with uncovering epistatic interactions is the volume of data to be analyzed; as the number of SNPs increases the number of potential interactions increases exponentially [9], known as the 'curse of dimensionality'. The potential complexity of such interactions supports the use of machine learning and data mining techniques.

Machine learning (ML) approaches employ algorithms to 'learn' from training data sets to solve problems and enable predictions about outcomes in other data based on patterns and rules learned. Classification and clustering analysis are two key roles of machine learning methods. Cluster analysis or clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval and bioinformatics. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy andtypological analysis [13].

It worth mentioning that the above said methods and algorithms can be implemented and run in most kinds of machines and computerized information systems (IS). For instance in [14] a novel approach, the profile distance method, has been presented to support the IS selection problems.

There are several issues that need to be considered when developing ML methods for the identification of epistasis including: genetic heterogeneity (which may be common in complex diseases by [15]), the presence (or absence) of main effects and the number of SNPs involved in the interactions.

Useful Machine Learning Methods in Bioinformatics: The most challenging problems which biologists and computer scientists face with in the field of bioinformatics are, including:

- Sequenceanalysis
- Genome annotation,
- Analysis of gene expression
- Analysis of mutations in cancer
- Protein structure prediction
- Comparative genomics
- Modeling biological systems
- High-throughput image analysis
- Protein-protein docking

A range of machine learning methods have been developed over the past two decades with the aim of uncovering computational biology problems implicated in common complex diseases. Here we discuss some approaches that have been used to detect epistasis, namely multifactor-dimensionality reduction (MDR), artificial neural networks (ANNs), random forest (RF) and support vector machines (SVMs).

Artificial Neural Network (ANNs): ANNs were originally developed to model neurons but are now regularly used for machine learning and data mining in a wide range of fields [16,17] with 'feed-forward/back-propagation' networks being the most common [18].

In essence, ANN is a system modelled on the human brain usuallyconsists of an input layer, some hidden layers and an output layer. The back-propagation algorithm was used for training of all ANN models [19].They have excellent power for performing pattern recognition and classification [15] and are capable of dealing with voluminous data [17]. AnANN resembling a directed graph where the nodes represent genetic elements (SNPs) and the arcs are the connections (interactions) between the elements, has been developed for genetic applications [17]. The nodes are arranged into layers. One or more nodes reside in the input layer and receive the information to be processed by the NN. The input layer links to multiple nodes in a hidden layer (of which there may be several) via arcs.

Finally, there is an output node. Each arc is assigned a weight which, initially, is chosen randomly, but through training the network on test data, weights are adjusted to minimize the error rate common [18].

The target of the NN is the recognition of corresponding patterns in real data, based on patterns observed in test data and for predictions about patterns not seen before through recognizing sub-patterns and correlations in the data [18]. To uncover genetic loci potentially involved in epistatic interactions Nns are trained using known genotypes as inputs and known phenotypes as outputs and the development of the internal weighting structure is of particular importance. The internal weight structure of the network can be analyzed after training todetermine the effect of each locus on the resulting phenotype [18]. NN applications to disease data have shown variable success. [17] Suggest that this may be due to the use of sub-optimal NN architecture. Exhaustive search of all possible architectures to find the optimal structure is infeasible and so one solution is to optimize architecture with ML algorithms. Examples of such algorithms include the Genetic Programming optimized NN (GPNN) [20,21] and Grammatical Evolution NN (GENN) [17]using genetic programming (GP) or grammatical evolution (GE) respectively to optimize a NN.

Genetic programming aims to 'evolve' computer programs to solve complex problems [22]. First, an initial population of randomly generated computer programs is produced. Each program is run on a problem and assigned a fitness value based on its performance. The best programs are chosen to go forward for 'reproduction' following the 'survival of the fittest' principle. Some programs are taken into the next generation unaltered, while others undergo 'crossover' in which new programs are created from combinations of components of the original programs. This procedure is repeated for a number of generations to find the optimal program [23].

GE is a variation of and improvement on, GP, with more flexibility. GE uses populations consisting of linear genomes which constitute individuals. Each genome is divided into codons which are translated into phenotypes (the NN) by the grammar. In a similar way to GP, the resulting phenotypes can be tested for fitness and subsequent generations produced to find the optimal model.

GPNN has higher power to detect gene-gene interactions in the presence of non-functional SNPs than the more traditional Back Propagation NN (BPNN) (Ritchie et al., 2003) while power comparisons have shown that GENN consistently outperforms GPNN [16,17] NNs can screen out loci that do not affect the phenotype, thus reducing the number of genetic locus combinations to be tested [18].Network approaches can also be used to identify genetic interactions through exhaustive enumeration of all possible pairwise interactions; however, this approach only searches for SNPs with strong pairwise interactions so may over- look SNPs with higher order interactions.Genetic heterogeneity, polygenic inheritance, high phenocopy rates and incomplete penetrance are problematic in the search for epistasis. Some of the characteristics of NN methods render them capable of addressing these difficulties in patterns.

Support Vector Machine (SVM): The SVM algorithm was first invented by VladimirVapnik in 1963 and then generalized for a non-linear state in 1995 by Corinna Cortes and Vapnik. The extended SVM algorithm by Vapnik was based on statistical learning theory and one of the most successful algorithms for classification.SVM encounters with classification problems by searching for hyper-planes in properties space and maximizing sample margins when test samples are separable. Support vector machines are one of the most popular classification algorithms in machine learning literature which can identify linear and non-linear decision ranges in test data and train data accurately. Basically, learning algorithms for SVM utilize for solving quadratic optimization equation.

The base of SVM classifier is linear data classification and tries to select lines with higher reliability margins. Problem solving is done through finding optimized line for data by QP methods which are well-known methods in solving constrained problems. Before linear division we map data to a very higher dimension space by ‘phi’ function for better machine’s classification of high complexity data. For solving high dimensionality problems with these methods we use Lagrange duple theorem for changing minimization problem to its duple form which takes us to a higher dimension space instead of ‘phi’ complex function. There is a more simple function called Kernel Function which is vector multiple of ‘phi’ function.

SVMs are classification techniques which are potentially as powerful as ANNs [24]. In the development of a supervised learning approach the actual outcome of the (training) data is given and similar patterns are searched for during testing [25]. In its simplest form, a SVM is focused on identifying a linear separator to divide data points of two classes and is thus a non-probabilistic binary linear classifier. Furthermore, using kernel functions, non-linear separators can be established by modifying the input space. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. SVMs have shown excellent power to detect epistasis in both simulated and real datasets[15];Listgartenin [25] identified variants in a number of genes associated with breast cancer risk. A quadratic kernel was used and the authors showed that multiple SNP sites from several genes at distant parts of the genome were better at identifying breast cancer patients than single SNPs. When compared to MDR this approach provides more interpretable output; however, unlike MDR, SVMs cannot cope well with missing data. [15] Employed an SVM approach which was combined with search algorithms to produce four different models to detect epistasis, in the absence of genetic heterogeneity.

Sparse SVMs [26] have been developed to select variables for inclusion in the model as a preprocessing step. This technique aims to reduce the instabilities in SVM results that arise from small changes in training/validation data. Such an approach might be usefully applied to the study of epistasis.

Random Forest (RF): Random Forest algorithm is a family member of classification methods based of several decision trees. The main characteristic of this group of classifiers is that their components grow like a tree in a random mode. Although this idea was extracted and implemented in 90’s, the formal definition and use of ‘Random Forest’ was first presented in an article in 2001 by Breiman in [27] which used two randomization rules: Bagging and random feature selection (RFS) as follows:

RF is a classifier consists of a set of tree-structural classifiers $\{h(x,\Phi_k) , k = 1, . . . ,L\}$ which $\{\Phi_k\}$ is independent and distributed random vectors and each $h(x,\Phi_k)$ tree has a valuable output for each input x. this selection is usually based on some degree of impurity which is a measure for determination of current nodes to partition to several child node. RFS chooses for each subset a set of properties randomly.

RF is a type of high-dimensional non-parametric predictive model composed of a collection of classification or regression trees [28] generated from random vectors [27]. Each tree of a RF is grown from a training set (or bootstrap sample) from the original data using random feature selection and trees are grown to their full extent without pruning. The bootstrap sample of size n is produced from the original sample, also size n, with variables chosen with replacement. Thus some variables will be chosen multiple times while others will not be chosen at all [28]. The best split at each node in each tree is chosen from a random subset of the predictor variables [29]. The so-called ‘out-of-bag’ (OOB) estimates of prediction error are then generated from the observations that are not chosen in the bootstrap sample (often up to one third of cases are not included). The RF algorithm is an effective prediction tool with the potential to uncover interactions among genes that do not exhibit strong main effects [22], however, it has been suggested that their ability to detect interactions actually depends on the presence of main effects, no matter how weak [30]. Thus, this approach may lack power to uncover those interactions that occur in the absence of any main effects. A recent study used the RF approach to uncover interacting SNPs contributing to rheumatoid arthritis, but no significant interactions were found that could be replicated in a follow-up cohort. Power calculations have further indicated that this method will only detect those interactions with a large effect size [31]. However, an advantage of RFs is that they do not ‘overfit’ the data and, as the number of trees in the RF increases, the prediction error converges to a limiting value [28].

An importance score is provided for each variable in a RF [29] rendering it capable of identifying SNPs predictive of a phenotype. This has prompted suggestions that RFs could be used to highlight significant SNPs for analysis with other methods [28]. However, this would conflict with the suggestion that RFs are useful tools to uncover genetic epistasis since the detection of interactions between variables is more important than the effect of single SNPs ondisease status. A further downside of the RF method is that, although it has shown considerable promise in low-dimensional data (100 SNPs and 10,000 observations), it has not been successfully applied to GWAS data [30].

Multifactor Dimensionality Reduction (MDR): MDR was one of the first ML methods developed to detect and characterize gene-gene interactions [32,33]. The MDR method was first suggested by Ritchie in 2001 and Moore and William in 2002 and then implemented by [34]. In the first stage of MDR, n genetic factors (e.g. SNPs) are selected from the entire set of factors. All possible multifactor (SNP genotype) combinations are represented in cells in n-dimensional space and each cell is assigned a case-control ratio.

Multilocus genotypic predictors are thus reduced from n dimensions to one dimension by classifying each cell as either low-risk or high-risk, based on a threshold value of cases-to-controls [32,35]. Following classification cross-validation is carried out to estimate the prediction error of each model by splitting the data into a training set consisting of 90% of the data and a testing set of the remaining 10%. A model is developed based on the classification of genotypes in the training set which is used to predict disease status of genotypes in the test set. The cross-validation process is repeated 10 times and the prediction error is averaged [32]. MDR modeling can thus be applied to real disease data to search for epistasis and any predictors designated as 'high-risk' are, therefore, potentially disease-related. This approach was evaluated using a sporadic breast cancer data set [32]. A statistically significant high-order interaction was detected amongst four polymorph- isms in the absence of any significant main effects, one of the earliest reports of such an interaction associated with a common multifactorial disease. The power of MDR was found to be robust to the presence of 5% genotyping error, 5% missing data and a combination of the two for a number of different two-locus epistasis models.

Some of the advantages of using MDR for the discovery of epistasis include:

- The model-free approach, invaluable for diseases such as sporadic breast cancer for which the mode of inheritance is unknown and likely to be complex.
- The capability of MDR for detecting and characterizing multiple genetic loci simultaneously and, through the use of cross-validation, minimizing the false-positive rate.
- The number of interaction terms does not grow exponentially as each new variable is added [32].

However, some disadvantages associated with this method impact upon its reliability as a predictor of disease-genotype interactions. In the presence of a high (50%) phenocopy-genetic heterogeneity rate, power is greatly compromised [34] supporting the need for refinements to effectively deal with genetic heterogeneity in complex trait data. The resulting models can be difficult to interpret [32], although genotypes are classified as 'high-risk' or 'low-risk' there is no quantitative assessment of how high or low risk they are, thus it is difficult to determine which of the putative interactions are most likely to be disease-related and warrant further investigation. MDR has only been successful when applied to a small number of SNPs in certain genes of (known) interest [32,33,36,37]. The MDR approach alone is not directly applicable to GWAS data, given the huge number of interactions to be assessed; however, using a filter algorithm to isolate a subset of potentially interesting SNPs for MDR analysis can overcome this limitation. Finally, MDR has a high false positive and negative error rate when the case and control ratio in a genotype combination is closely similar to that in the whole data set [36].

Comparison Study of the Four Methods: In this section we compare the four above said machine learning algorithms and survey some of their advantages and disadvantages in their different applications. First of all we should take note that there is no perfect machine learning method.Different methods have different results and applications for different problems and this is our duty to choose the best one for our problem. The more chosen techniques better suits the problem;the final results are more accurate. For instance in image processing problems ANNs method better works, while in problems like text processing and extraction, intrusion detection in security issues and medical sciences the SVM method has a higher amount of speed and accuracy. In gene selection problems RF better works in comparison with SVM.

MDR: As it is clear from its name, the multifactor dimensionality reduction methods applies on problems with 2 or 3 dimension of size which are computationally feasible. If the computationally facilities are limited we can use another method like Genetic algorithm. The MDR is a powerful method in gene-gene interaction detection problems; although there are some challenges in genotypes errors, lack of data, phenocopy and genetic inheritance problems. MDR is a suitable substitution method for parametric techniques like logical regression. One of the disadvantages of MDR method which is the main challenge of most of machine learning algorithms is 'overfitting problem'; even though this can be solved by implementation of MDR in a cross-validation framework for assessment of ability to predict models.

ANNs:
The Pros of a Neural Network Are, Including:

- Operating specific tasks that a linear program cannot,
- Parallel origin which helps keep continue working even there is a fault in one of its elements,
- Ability to learn which do not need to be programmed,
- Ability to implement on almost every application by no trouble,
- Simplicity of implementation,
- Flexibility if support of all data types, etc.

The Cons of a Neural Network Are, Including:

- The need of training and learning to run,
- Long processing time for big neural networks,
- More useful in cognitive science than in operational cases because they show patterns resemble to human's behavior
- Time series management in this method is more complex, etc.

SVM:
Svm Method's Positive Points Are, Including:

- One of the best methods in data classification and regression,
- More efficient than other methods like nearest neighbor, neural network, decision tree,
- A powerful theoretical background,
- Lack of problem in thenumber of parameters selection,
- Less tendency in 'overfitting',
- Less memory requirement to save predictive model,
- Production of more geometric readable and interpretable results,
- More useful for unsupervised learning,
- Unlike ANN method, the computational complexity is not depending on input space dimensions, etc.

SVM Method's Negative Points Are, Including:

- Constraints in Kernel selection method,
- Constraints in size and speed, especially in training and testing mode,
- Slowness in testing phase,
- High algorithmic complexity,
- High saving memory, etc.

RF:
Random Forest Method's Advantages Are, Including:

- Non-parametric, interpretable and optimize method,
- High accuracy in prediction of most types of data and applications,
- Including most of decision trees advantages for getting better results,
- Including a great amount of variables (continues, binary, indexed),
- Suitable for high dimension data modeling,
- Simplicity, accuracy, speed, robustness, etc.

Some of the random forest method's disadvantages are difficulty in data and models interpretation and 'overftting' on some of classification and regression datasets.

CONCLUSION

Information technologies have made numerous progresses [38-41] Several research have been conducted on various areas of IT [38-43] including bioinformatics. All in all, after reviewing the application of four machine learning algorithms in bioinformatics and its sub-fields like gene-gene interactions, we now have a better outlook on methods and techniques in order to better work with them and finally reach desired results. As an example, it has

Table 1: Machine learning methods comparison table

Method	Advantages	Disadvantages	Application
MDR	-High dimensions and factors problem solving -Suitable for non-additive gene-gene interactions detection -Suitable substitute for statistical parametric methods like regression	-Overfitting problem -High computational complexity	-Epistasis detection in breast cancer -High order identification of gene-gene interactions
ANNs	-Learns and does not need to be programmed -Simplicity of implementation -All data types support	- High processing time for big neural networks - More useful in cognitive science	- Pattern recognition in genetic heterogeneity and polygenic inheritance problem solving - Signal filtering in high phenocopy rates and incomplete penetrance
SVM	- Regression and classification problem solving - Powerful theoretical background - Low computational complexity - High algorithmic complexity	- Constraints in size and speed in both train and test mode - Constraints in memory space	- Diseases diagnosis and prediction - Epistasis detections
RF	- Non-parametric - Optimized - Accuracy - Simplicity of use - All types variables	- Difficulties in data interpretation - Overfitting problem	- Gene-gene interaction detection - SNPs interaction detection - Gene selection

been suggested that RF methods may be successful at dealing with certain types of heterogeneity, while some of the characteristics of ANNs render them capable of addressing genetic heterogeneity, polygenic inheritance, high phenocopy rates and incomplete penetrance. The SVM method as one of the best techniques in regression and classification problems works very well in applications like diseases diagnosis and prediction and epistasis detection in medical informatics field, although there are some stumbling blocks like constraints and limitations in speed, size and saving memory.

As far as application of machine learning techniques in bioinformatics is concerned, there is no perfect method to solve a biological problem; however, most of the times we better compare them with each other and then apply them in our problem. Given the increasingly voluminous genetic data now being produced by next generation sequencing studies and the emerging evidence that very large numbers of individually low risk variants underlie common diseases, the need for powerful ML models is more pressing than ever. It is evident that current methods require further development before successful application to these enormous data sets can be claimed and their outputs enhance understanding of the genetic epidemiology of disease or become useful in a clinical disease risk predictive setting.It worth mentioning that some other similar researches have been done on application of Machine learning methods in some specific field of biology [44].

REFERENCES

1. Fakruddin, Reaz Mohammad Mazumdar et al., 2013. Pyrosequencing-A Next Generation Sequencing Technology, World Applied Sciences Journal, 24 (12): 1558-1571, 2013 ISSN 1818-4952.
2. Muneer Ahmad, Azween Abdullah and Khalid Burraga, 2010. Optimal Nucleotides Range Estimation in Diffused Intron-exon Noise, World Applied Sciences Journal, 11(2): 178-183, 2010 ISSN 1818-4952.
3. Moore, J.H. and S.M. Williams, 2009. Epistasis and its implications for personal genetics, Am. J. Hum Genet, 85: 309-20.
4. Hirschhorn, J.N., 2009. Genomewide association studies-illuminating biologic pathways, N. Eng. J. Med., 360: 1699701.
5. Cordell, H.J., 2009. Detecting gene-gene interactions that underlie human diseases, Nat. Rev. Genet, 10: 392-404.
6. Hindorff, L.A., P. Sethupathy, H.A. Junkins, et al., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad Sci. USA, 106: 9362-7.
7. Stratton, M.R. and N. Rahman, 2008. The emerging landscape of breast cancer susceptibility, Nat. Genet, 40: 17-22.
8. Harlid, S., M.I.L. Ivarsson, S. Butt, et al., 2012. Combined effect of low-penetrant SNPs on breast cancer risk, Br. J. Cancer, 106: 389-96.
9. Moore, J.H.P. and M.D.P. Ritchie, 2004. The challenges of whole- genome approaches to common diseases, JAMA, 291: 1642-3.
10. Moore, J.H. and S.M. Williams, 2005.Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bio Essays, 27: 637-46.
11. Phillips, PC., 2008. Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet, 9: 855-67.
12. Cordell, H.J., 2002. Epistasis: what it means, what it doesn’t mean and statistical methods to detect it in humans. Hum. Mol. Genet, 11: 2463-8.

13. Bhuvana, B.P., 2014. Segmentation of Brain MRI Images by Using Modified Robust Fuzzy c Means Algorithm, World Applied Sciences Journal, 29(10): 1327-1332, 2014 ISSN 1818-4952.
14. Edward, W.N. Bernroider and Volker Stix, 2006. Profile distance method - a multi attribute decision making approach for information system investments, Decision Support Systems, 42: 988-998.
15. Chen, S.H., J. Sun, L. Dimitrov, et al., 2008. A support vector machine approach for detecting gene-gene interaction,Genet Epidemiol., 32: 152-67.
16. Motsinger, A., S. Dudek, L. Hahn, et al., 2006. Comparison of neural network optimization approaches for studies of human genetics, Lect Notes Comp Sci., 3907: 103-14.
17. Motsinger-Reif, A.A., S.M. Dudek, L.W. Hahn, et al., 2008. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology,Genet Epidemiol., 32: 325-40.
18. Lucek, P.R. and J.Ott, 1997. Neural network analysis of complex traits, Genet Epidemiol., 14: 1101-6.
19. Mahdavian, A., A. Banakar, A. Mohammadi, M. Beigi and B. Hosseinzadeh, 2012. Modelling of Shearing Energy of Canola Stem in Quasi-Static Compressive Loading Using Artificial Neural Network (ANN), Middle-East Journal of Scientific Research, 11(3): 374-381, 2012 ISSN 1990-9233.
20. Ritchie, M.D., A.A. Motsinger, W.S. Bush, et al., 2007. Genetic programming neural networks: a powerful bioinformatics tool for human genetics,Appl Soft Comput., 7: 471-9.
21. Koza, J.R and J.P. Rice, 1991. Genetic generation of both the weights and architecture for a neural network, Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference, 392: 397-404.
22. Moore, J.H., F.W. Asselbergs and S.M. Williams, 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics, 26: 445-55.
23. Ritchie, M., B. White, J. Parker, et al., 2003. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics, 4: 28.
24. Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach Learn, 20: 273-97.
25. Listgarten, J., S. Damaraju, B. Poulin, et al., 2004. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res., 10: 2725-37.
26. Bi, J., K. Bennett, M. Embrechts, et al., 2003. Dimensionality reduction via sparse support vector machines. J. Mach Learn Res., 3: 1229-43.
27. Breiman, L., 2001. Random forests. Mach Learn, 45: 5-32.
28. Bureau, A., J. Dupuis, K. Falls, et al., 2005. Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol., 28: 171-82.
29. Yoshida, M. and A. Koike, 2011. SNPInterForest: a new method for detecting epistatic interactions. BMC Bioinformatics, 12: 469.
30. Schwarz, D.F., I.R. Konig and A. Ziegler, 2010. On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics, 26: 1752-8.
31. Liu, C., H. Ackerman and J. Carulli, 2011. A genome-wide screen of gene-gene interactions for rheumatoid arthritis susceptibility. Human Genetics 129: 473-85.
32. Ritchie, M.D., L.W. Hahn, N. Roodi, et al., 2001. Multifactor- dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet, 69: 138-47.
33. Gui, J., A.S. Andrew, P. Andrews, et al., 2011. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. Ann. Hum. Genet, 75: 20-8.
34. Hahn, L.W., et al., 2003. Multifactor-dimensionality Reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics, 19: 376-382.
35. Ritchie, M.D., L.W. Hahn and J.H. Moore, 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy and genetic heterogeneity. Genet. Epidemiol., 24: 150-7.
36. Chung, Y., S.Y. Lee, R.C. Elston, et al., 2007. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics, 23: 71-6.
37. Lou, X.Y., G.B. Chen, L. Yan, et al., 2007. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am. J. Hum Genet, 80: 1125-37.
38. Ghapanchi, A.H., A. Aurum and F. Daneshgar, 2012a. The Impact of Process Effectiveness on User Interest in Contributing to the Project, Journal of Software, 7(1): 212-219.

39. Ghapanchi, A.H., M. Tavana, M.H. Khakbaz and G. Low., 2012b. A Methodology for Selecting Portfolios of IS/IT Projects with Interactions and Under Uncertainty. International Journal of Project Management, 30(7): 791-803.
40. Khakbaz, M.H., A.H. Ghapanchi and M. Tavana, 2010. A multi-criteria decision model for supplier selection in portfolios with interactions. International Services and Operations Management, 7(3): 351-377.
41. Ghapanchi. A.H. and A. Aurum, 2011. Measuring the Effectiveness of the Defect-Fixing Process in Open Source Software Projects, 44th Hawaii International Conference on System Sciences (HICCS), 4-7 January 2011, Hawaii, US.
42. Ghapanchi, A.H., M.H. Khakbaz and M.H. Jafarzadeh, 2008. An Application of Data Envelopment Analysis (DEA) for ERP System Selection: Case of a Petrochemical Company, International conference of information systems (ICIS), December 2008, France, Paris.
43. Ghapanchi, A.H. A. Aurum and G. Low, 2011. Creating a Measurement Taxonomy for the Success of Open Source Software Projects, First Monday, 16(8).
44. Badrinath, N. and G. Gopinath, 2014. A Survey on Utilization of the Machine Learning Algorithms for the Prediction of Erythemato Squamous Diseases, World Applied Sciences Journal, 29 (6): 752-757, ISSN 1818-4952.