Autism with Optimized Machine Learning Models

Subjects: Computer Science, Artificial Intelligence

Contributor: Heyam Al-Baity , Maraheb Alsuliman ,

Early diagnosis of autism is extremely beneficial for patients. Traditional diagnosis approaches have been unable to diagnose autism in a fast and accurate way; rather, there are multiple factors that can be related to identifying the autism disorder. The gene expression (GE) of individuals may be one of these factors, in addition to personal and behavioral characteristics (PBC). Machine learning (ML) based on PBC and GE data analytics emphasizes the need to develop accurate prediction models. The quality of prediction relies on the accuracy of the ML model. To improve the accuracy of prediction, optimized feature selection algorithms are applied to solve the high dimensionality problem of the datasets used. Comparing different optimized feature selection methods using bio-inspired algorithms over different types of data can allow for the most accurate model to be identified.

autism spectrum disorder (ASD) big data bioinformatics machine learning classification bio-inspired algorithms

1. Introduction

Autism spectrum disorder (ASD) is a neurological developmental disorder. It affects how people connect and interact with others and how they behave and learn [1]. The symptoms and signs appear when a child is very young. It is a lifelong condition and cannot be cured. Today, ASD is one of the fastest-growing developmental disorders, resulting in many problems, such as school problems related to successful learning, psychological stress within the family, and social isolation. However, early diagnosis can help the family take preliminary and effective steps to ensure the normal life of the patient. It can help providers of healthcare and families of patients by affording the effective therapy and treatment required, thereby reducing the costs associated with delayed diagnosis. On the other hand, many factors can be used to detect ASD cases, including personal and behavioral characteristics, genetic, brain images, and family history. Notwithstanding its genetic causes, ASD is mainly diagnosed utilizing personal and behavioral indicators that are tested in traditional clinical examinations by different specialists during regular visits. However, these traditional clinical methods, which primarily depend on the clinician, are time consuming and cumbersome. Currently, with computer power and big data generated by hospitals such as clinical data, gene expression profiles, and medical imaging, ASD can be automatically predicted and diagnosed in its early stages by using predictive models that use big data sets with ML algorithms, which can improve the life quality of patients and families as well as reduce the financial costs.

The personal and behavioral characteristics (PBC) and the gene expression (GE) data are the most available and valuable resources for machine learning (ML) algorithms seeking to discover new and hidden patterns of data to help in ASD prediction, thus helping families to take early steps for treatment. Nevertheless, the high dimensionality of these data makes the prediction process challenging. The feature selection (FS) mechanism can help in reducing the high dimensionality of such datasets, increasing the speed of the classification process, decreasing the cost, and improving the accuracy of the prediction models by selecting the most effective features.

Feature selection algorithms [2] aim to choose the most significant features to solve the prediction problems. In general, there are three common types of FS algorithm: filter, wrapper, and hybrid. Due to the potential benefits that can be achieved from automatic ASD classification, research in this field has recently gained much attention. Several methods have been proposed to solve the problem of predicting ASD. However, it is still an open problem and further improvement can be achieved.

Bio-inspired algorithms are one of the techniques that can be integrated into the wrapper feature selection method to search globally for the optimal feature subset and improve prediction accuracy [3]. They can be classified as a type of Nature-inspired Computation algorithms that rely on the inspiration of the biological evolution of nature to provide new optimization techniques. A number of researchers have adopted bio-inspired techniques for dealing with the high dimensionality of features, and they have shown high results in improving the diagnosis process of many diseases such as cancer [4]. However, there are few studies in research on ASD prediction using optimized FS algorithms and further investigation in this field is needed. To the best of knowledge, this is the first study to deal with this problem using four bio-inspired algorithms (GWO, FPA, BA, and ABC). In addition, this is the first study that employed the CNN deep learning approach for ASD GE and PBC datasets.

2. Background

2.1. Personal and Behavioral Characteristics (PBC)

At clinical diagnosis, clinicians use questionnaires and behavioral observation to collect personal and behavioral information based on the Manual of Mental Disorders (DSM-5) criteria, which include two main symptoms. The first symptom is a chronic deficiency in social communication and social engagement through various contexts. The second symptom is minimal and repeated behavior patterns, desires, and behaviors. Personal and behavioral data generally include tens of attributes (high dimensionality) that can be classified into personal information (such as age, ethnicity, and born with jaundice) and behavioral screening questions (such as “Do ASD patients often hear small sounds when others do not?” or “Is it difficult to hold the attention of ASD patients?”) [5].

2.2. Gene Expression Profile (GE)

Gene expression is the mechanism by which the information stored in the gene is used to guide the assembly of the protein molecules. DNA microarray technology has become an effective way of tracking gene expression levels within the organism for biologists [6]. This technique helps researchers to assess the expression levels of a set of

genes. Gene expression data usually comprise a wide range of genes and a small number of samples (high dimensionality). In medical fields, microarray technology is most widely used to find out what reasons and how to cure illnesses. Researchers have found that often the cause of some diseases, such as ASD, may be DNA mutations. It is well known that certain disorders are caused by the mutation of certain known genes. There is however, no particular form of mutation that causes all disorders. Therefore, the microarray gene expression analysis is used to identify and diagnose common genes mutations. Analysis of GE data is the method of identifying the helpful genes in the diagnosis.

2.3. Classification Algorithms

Researchers used four different classification algorithms to analyze the datasets: support vector machine (SVM), decision tree (DT), Naïve Bayes (NB), and k-nearest neighbor (KNN) algorithms.

SVM [7] is one of the classification algorithms, and classifies two data types: linear and nonlinear.

First, the training dataset is converted into a higher dimension using nonlinear mapping. Next, it looks for linear separating hyperplanes (which are decision boundaries that help classify the data points) in the new dimension and splits the data based on the class. The optimal hyperplane [7] separates data points into classes that can be specified based on margin and support vectors. Support vectors are identified as the closest points of each class to the margin line. The NB classifier is based on Bayes' theorem and is a probabilistic classifier. The presumption of conditional independence underpins this classifier. This implies that the values of the attributes for each class mark are effectively conditionally independent of one another. Despite this basic assumption, Naïve Bayes has been successfully applied to a variety of real-world data circumstances [8]. KNN is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. It is based on the similarity measure to classify the new cases by calculating the distance measured from the trained available cases. In DT, the data are visualized using a tree structure, which is represented as sequences and consequences using the decision tree. The root node is at the top of the tree, while the internal nodes are where the attributes are tested. The result of the test is represented by the "branch". Finally, leaf nodes are nodes that have no further branching and indicate the class label of all previous decisions.

2.4. Feature Selection (FS)

Feature selection, as a data preprocessing technique, has been shown to be effective and efficient in preparing high-dimensional data for ML problems. The objectives of the selection of features include the development of simpler and more comprehensible models, the improvement of ML efficiency, and the preparation of clean and understandable data. The recent proliferation of large data has posed some major challenges and opportunities for feature selection algorithms [9]. The most common feature selection techniques are as follows: The filter approach, where the typical features are ranked via specific criteria. Features are then identified with the highest ratings then used as inputs for the wrapping or classification process [8][10]. On the other hand, the definition of the wrapper method requires the use of learning strategies to choose the optimum function subset to be used in the

classification process. Usually, the wrapper method uses nature-inspired computational algorithms (NICs) to direct the search process by choosing the optimum feature subsets. The third approach is hybrid, which uses both filter and wrapper approaches. Based on [11], feature selection is a difficult task due to the need for searching over a large space, which is impossible in some applications that have large features and small samples. This problem can be solved using NIC algorithms that are able to search globally and can be utilized to solve the feature selection problem.

2.5. Nature-Inspired Computation (NIC)

NIC [12] refers to algorithms that imitate or optimize the behavior of natural and biological systems to solve problems in order to overcome or optimize the limitations of certain algorithms. All these algorithms share two characteristics: natural phenomena are replicated and modelled. NIC algorithms can be categorized into four types: swarm intelligence, bio-inspired, physics and chemistry, and other algorithms [13].

2.6. Bio-Inspired Algorithms

This is an emerging approach, focused on the inspiration of the biological evolution of nature, to develop new competing techniques. Bio-inspired optimization algorithms have demonstrated greater performance in a variety of disciplines, including disease diagnosis, by using the wrapper technique to high-dimensional datasets for feature selection. Algorithms for bio-inspired optimization are usually classified into three categories. Some of the well-known bio-inspired algorithms are described in the following section and are shown in **Figure 1**.

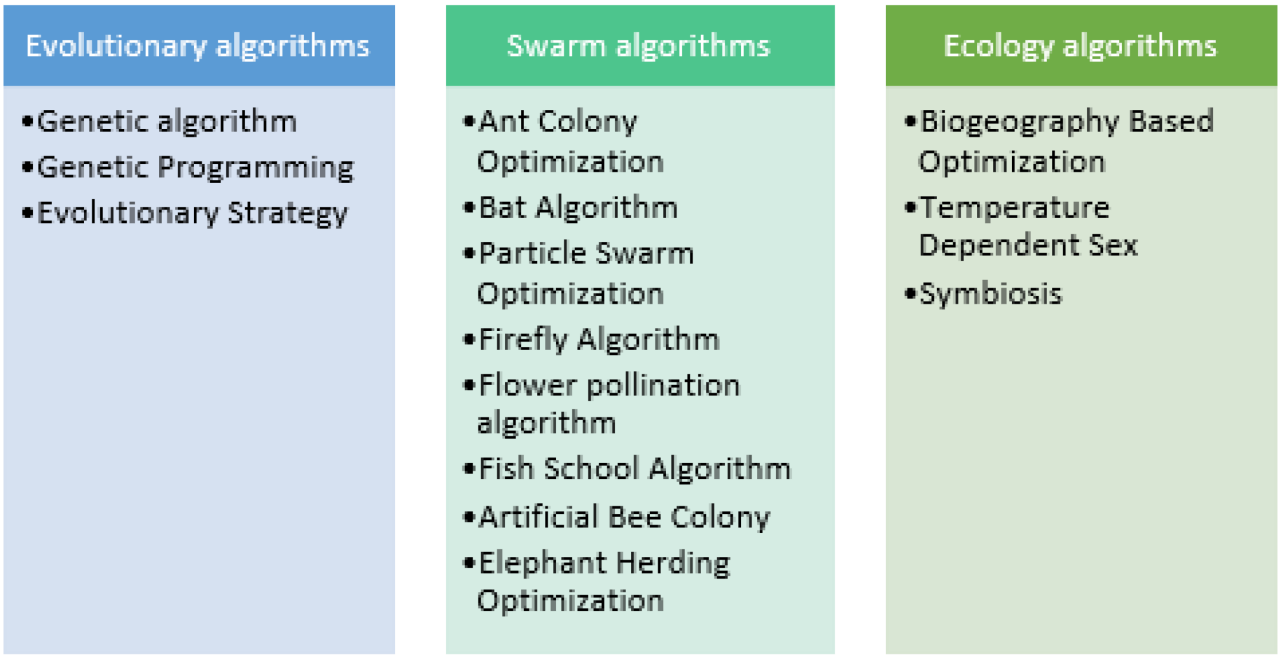

Figure 1. Bio-inspired algorithms.

2.7. Grey Wolf Optimization (GWO)

GWO algorithm is a recent algorithm proposed in 2014 [14]. This algorithm mimics the social behavior of grey wolves while searching and hunting for the prey. Normally, the wolves live in a pack with a group size of 5 to 12. The wolves are guided by three leaders, namely, alpha, beta, and delta wolves. The alpha wolf is responsible for making the decision, the beta wolf helps the alpha wolf in decision-making or pack activity, while the delta wolf submits to the alpha and beta, and dominates the omega wolves.

2.8. Bat Algorithms (BA)

This is one of the newest micro-bat algorithms, naturally inspired, utilizing echolocation behavior to locate their prey. To measure size, echolocation is used by bats. Therefore, in order to pick the booty (solution), they randomly migrate to particular locations at a given velocity and at a set frequency. Among the best solutions, the solution is selected and created through the use of random walking [15].

2.9. Flower Pollination Algorithms (FPA)

The flower pollination algorithm, one of the newest optimization algorithms, is inspired by the action of flower pollination. Crop pollination strategies in nature include two primary types: cross-pollination and self-pollination [16]. Some birds act as global pollinators in cross-pollination, passing pollen to the flowers of more distant plants. On the other hand, pollen is spread by the wind and only among adjacent flowers in the same plant during self-pollination. The FPA is therefore established by mapping the two types of cross-pollination and self-pollination into global pollination operators and local pollination operators. Due to the merits of fundamental principles, few parameters, and ease of operation, the FPA has attracted considerable interest.

2.10. Artificial Bee Colony (ABC)

This is an organic algorithm inspired essentially by the behavior of bees in the search for good sources of food. The ABC algorithm consists of three classes of bees: employed bees, onlooker bees, and scout bees. The employed bees find a source of food as well as exchange information of the source of food with the employed bees in the hive who are waiting for dancing. The onlooker bees choose a good source of food from the discovered food. The bees that choose the food sources at random are known as scout bees. Any bees that do not change their food source become scout bees [17].

References

1. Hirvikoski, T.; Mittendorfer-Rutz, E.; Boman, M.; Larsson, H.; Lichtenstein, P.; Bölte, S. Premature mortality in autism spectrum disorder. Br. J. Psychiatry 2016, 208, 232–238.
2. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Feature selection for high-dimensional data. Prog. Artif. Intell. 2016, 5, 65–75.

3. Vaishali, R.; Sasikala, R. A machine learning based approach to classify autism with optimum behavior sets. Int. J. Eng. Technol. 2018, 7, 1–6.
4. Al-Baity, H.H.; Al-Mutlaq, N. A New Optimized Wrapper Gene Selection Method for Breast Cancer Prediction. Comput. Mater. Contin. 2021, 67, 3089–3106.
5. Erkan, U.; Thanh, D. Autism Spectrum Disorder Detection with Machine Learning Methods. Curr. Psychiatry Rev. 2019, 15, 297–308.
6. Raza, K. Analysis of Microarray Data Using Artificial Intelligence Based Techniques; IGI Global: Hershey, PA, USA, 2016; pp. 216–239.
7. Suthaharan, S. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning; Springer: New York, NY, USA, 2015; p. 36.
8. Almugren, N.; Alshamlan, H. A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification. IEEE Access 2019, 7, 78533–78548.
9. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. ACM Comput. Surv. 2017, 50, 94:1–94:45.
10. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; de Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. 2012, 9, 1106–1119.
11. Sheikhpour, R.; Sarram, M.A.; Sheikhpour, R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Appl. Soft Comput. 2016, 40, 113–131.
12. Fan, X.; Sayers, W.; Zhang, S.; Han, Z.; Ren, L.; Chizari, H. Review and Classification of Bio-inspired Algorithms and Their Applications. J. Bionic Eng. 2020, 17, 611–631.
13. Fister, I., Jr.; Yang, X.-S.; Fister, I.; Brest, J.; Fister, D. A Brief Review of Nature-Inspired Algorithms for Optimization. Elektrotehniski Vestn./Electrotech. Rev. 2013, 80, 116–122.
14. Applying Grey Wolf Optimizer-Based Decision Tree Classifer for Cancer Classification on Gene Expression Data | IEEE Conference Publication | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/7365818 (accessed on 17 April 2021).
15. Yang, X.-S. A New Metaheuristic Bat-Inspired Algorithm. In Nature Inspired Cooperative Strategies for Optimization (NICSO 2010); González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 65–74.
16. Dankolo, N.; Radzi, N.; Sallehuddin, R.; Mustaffa, N. Hybrid Flower Pollination Algorithm and Support Vector Machine for Breast Cancer Classification. J. Technol. Manag. Bus. 2018, 5, 1.

17. A Simple and Efficient Artificial Bee Colony Algorithm. Math. Probl. Eng. 2013, 2013, 526315. Available online: https://www.hindawi.com/jour-nals/mpe/2013/526315/ (accessed on 3 December 2020).