IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 August 2024, accepted 13 September 2024, date of publication 18 September 2024, date of current version 1 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3463506

RESEARCH ARTICLE

Optimizing Business Intelligence Classification Rule Mining Using Quantum-Inspired Genetic Algorithm

YOUCEF ISLEM ADNANE AND MOUNIRA ZERARI
MISC Laboratory, NTIC Faculty, University Abdelhamid Mehri Constantine 2, 25000 Constantine, Algeria
Corresponding author: Youcef Islem Adnane (youcef.adnane@univ-constantine2.dz)

ABSTRACT This research introduces a novel approach to enhance knowledge discovery through the construction of classification rules for improved event prediction within the domain of business intelligence. One of the applications of the solution can be the improvement of large-scale advertising campaign forecasting. In order to solve the existing problems of data imbalance and high dimensionality in practical applications, we present a new approach of quantum-inspired genetic algorithm (QIGA). Our methodology employs a two-phase strategy: the first phase utilizes QIGA for fast attribute selection while the second phase enhances the manner in which categorical data is represented to enhance rule interpretability. This combined approach is very efficient in producing accurate and usable classification rules without requiring a lot of data pre-processing. Experimental evaluation on benchmark datasets demonstrates the superiority of our proposed method in terms of precision, reliability, and efficiency when compared to traditional approaches. Thus, the proposed algorithm enhances the state-of-art in knowledge discovery by providing a practical and scalable solution to the problem of handling imbalanced data and reducing the effect of high dimensionality, which allows for improving the accuracy of event prediction. Thus, the findings of this research can be applied to various fields other than advertising including fraud detection, customer churn, and medical diagnosis.

INDEX TERMS Business intelligence, classification rule, data mining, genetic algorithm, quantum-inspired algorithm.

I. INTRODUCTION

Business Intelligence (BI) assumes a pivotal role in addressing the myriad of challenges faced by businesses, particularly those operating in data-intensive sectors like banking [1]. The exponential growth in data volume necessitates efficient structuring and processing, making BI an indispensable solution. It encompasses a diverse range of practices, methodologies and applications that enable seamless access to information and empower its analysis, thereby facilitating the decision-making processes.

BI operates as a multifaceted framework that facilitates the implementation of strategies capable of addressing both short-term and long-term business needs [2]. Its scope extends across various domains, encompassing the functional aspects of data communication between services, servers, and data warehouses, as well as including knowledge discovery, forecasting, and decision support for enhancing performance outcomes [2]. By leveraging data-driven insights, BI serves as a comprehensive apparatus for driving business success.

Tailoring solutions to specific challenges represents a key aspect of BI. Whether it involves the development of projects aligned with the organization's overarching vision or the implementation of performance metrics to optimize business operations [3], BI ensures meticulous collection of relevant data through the Extract, Transform, Load (ETL) process [4]. This data then undergoes rigorous modeling within Data Warehouses, followed by comprehensive analysis and visualization [4], facilitating informed decision-making processes.

In this paper, our primary focus revolves around Knowledge Discovery through the construction of classification rules, which plays a crucial role in the field of data analysis

The associate editor coordinating the review of this manuscript and approving it for publication was Nadeem Iqbal.

within Business Intelligence (BI) [5]. The objective is to improve event prediction, more specifically the determination of the success of the large-scale advertising campaign or the identification of target customers for the campaign to yield the highest returns. Numerous approaches have been explored in this domain, falling into broad categories such as linear regression and artificial neural networks [6], decision and classification trees [7] including C4.5 [8] and classification and regression tree algorithm (CART) [9], bio-inspired algorithms like ant colony optimization (ACO) [10], [11], and evolutionary programming models and their variants [12], [13]. Also, the recent development in decision and classification has brought new techniques into the limelight including Time Delay Neural Network algorithm (TDNN) by Tripathi et al. [14] and Decision Table and Naíve Bayes algorithm (DTNB) by Mattiev and Kavsek [15]. In addition, the field of bio-inspired algorithms has seen new algorithms such as LVQ+varPSO wich is an hybridation of a varient of the Particle Swarm Optimization (PSO) and Learning vector quantization (LVQ) introduced by Santana et al. [16], (ASOA) Adaptive Seagull Optimization Algorithm introduced by Bangyal et al. [17] and other contributions by Anwaar et al. [18]. Finally, the integration of two or more approaches has been developed, for instance, a new varient of Decision Tree algorithm named (CDBT) by Weng and Huang [19] and (EFO-ANN) Artificial Neural Networks (ANN) combined with Electromagnetic Field Optimization (EFO) by Guo et al. [20].

However, one major drawback of these approaches is the lack of focus on the reliability aspect, particularly when working with the imbalanced data sets. From the papers which we discussed, namely Lee and Chen [6], Kesavaraj et al. [8], and Seidlova [11], it is seen that all the datasets used have a significantly larger majority class than the minority class. For instance, in Lee and Chen [6], the dataset used was the Taiwan Bank dataset with 510 cases, 459 of which were in the same class. Likewise, in Kesavaraj et al. [8], the AP liver dataset is comprised of 416 cases of liver disease patients while 167 patients are without the disease. In addition, in Seidlova [11], it is observed that the class imbalance ratio is above 75% in two of the three datasets, namely Adult and Bank. Notably, none of these papers mention preprocessing techniques or provide measures to assess how well their approaches handle this constraint, highlighting a significant gap in addressing the reliability aspect of predictions.

The notable strength of our proposed approach resides in the configuration of our algorithm, enabling the effective handling of the inherent imbalance constraint without the need for preprocessing techniques. Our algorithm's configuration is designed to strike a balance between precision optimization and maintaining reliability, with our quantum-inspired approach proving to be highly advantageous in this regard. By addressing the challenge of dimensionality within the optimization process, our approach facilitates enhanced accuracy in a more efficient and expeditious manner compared to conventional methods. Consequently, this mitigates the risk of overfitting and yields an algorithm that is both more reliable and robust in its predictive capabilities.

The remainder of the article is structured as follows. Section two, relates to the basic concepts to be acquired to be able to understand the work and be familiarized with both the data part including the classification rules and their specificities, as well as the fundamentals of the genetic algorithm in its quantum variant. Section three, contains everything related to our own work. It is organized in sub-headings that are, our motivation, our solution as well as its adaptation to the problem. The results obtained and the conclusion are in the fourth section.

II. BASIC CONCEPTS

In this section we will give an overview of current and ongoing works on classification rules and quantum-inspired genetic algorithms. We will define some basic concepts of the field as well as the methodology chosen in our solution.

A. CLASSIFICATION RULES

In the process of Knowledge Discovery, classification is a task of Data-mining which consists of seeking the criteria that make it possible to identify classes in a population of data [21].

1) DEFINITION AND PROPERTIES

Classification rules are assertions that link data relationships to class outcomes [22]. They emerge from an induction process using supervised training, distinct from association rules [23]. These rules follow a structured format: IF a certain condition holds, THEN assign a specific class [8], [24]. The antecedent (conditional part) is comprised of terms connected by logical operators, while the consequence represents the predicted class [8], [24].

When designing classification rule algorithms, several key properties are important. Accuracy, the most common measure, reflects how well the algorithm predicts classifications [25], but it can be misleading in imbalanced datasets [26]. To address this, reliability measures like F-measure or Kappa score are recommended [26]. Additionally, comprehensibility, often measured by rule length, ensures human users can easily understand the classification logic [27]. Finally, interestingness, encourages the algorithm to generate diverse, surprising, and non-repetitive rules while maintaining acceptable reliability [25]. These properties together guide the development of effective classification rule algorithms.

2) RELATED WORKS

In the work of Romana Seidlova et al. [11], to respond to the task of classification in the field of banking and marketing, an ACO-type approach was developed. The latter consists of an improvement of the Ant-Miner algorithm initiated by Parpinelli, Lopes and Freitas [10] and whose main characteristics are based on an agent-based system which simulates the movement behavior of an ant colony. The paths taken are considered as potential solutions (classification rules) [10]. The optimization provided by Seidlova et al. [11] on the parameters of the number of ants as well as the number of uncovered cases has made it possible to obtain excellent results in terms of accuracy. However, referring to a single performance metric is insufficient to highlight the full potential of a solution [26]. There are indeed other aspects on which the algorithm can bring an equally crucial result and which have not been put forward in this case.

Weng and Huang [19], proposed a hybrid machine learning framework that integrated feature selection, instance selection, and decision tree induction for credit approval. The model achieved better results than the classical methods of machine learning while actually reducing computational efforts. Inherent gain ratio usage in feature and instance selection improved the accuracy of the model. Though the hybrid systems have one prime flaw: its added complexity could be a major issue that requires higher computational power or training time. This creates the dilemma of model sophistication at the expense of computational efficiency, which may not work well in real-time or resource-constrained settings.

Guo et al. [20], describes an ensemble approach for bank telemarketing success rate prediction using ANNs trained by metaheuristic algorithms. In this regard, they compared four models: EFO, FSA, HSA, and SSD. The accuracy obtained by the EFO-ANN model was approximately 0.80, while the other models followed with AUCs of 0.7714, 0.7663, and 0.7160 for the FSA-ANN, SSD-ANN, and HSA-ANN, respectively. The decision-making process that lies underlying the suggested model is too complex to be understood by any banking professional, and this might hamper its practical implementation and acceptance in real life.

A comprehensive analysis of machine learning techniques for credit score classification was conducted by Tripathi et al. [14]. They employed nine feature selection methods and sixteen classification algorithms across seven benchmark datasets, including Bank Marketing, Australian, and German, the researchers demonstrated the pivotal role of feature selection in enhancing model performance. Their findings highlight the superior accuracy (92.3%) of Recursive Feature Elimination (RFE) coupled with Support Vector Machines (SVM). Moreover, a novel hybrid approach integrating Particle Swarm Optimization (PSO) with Neural Networks (NN) exhibited robust and accurate classification. While the study provides valuable insights, its exclusive reliance on benchmark datasets might limit the generalization of results to real-world credit scoring scenarios, necessitating further exploration of diverse datasets.

Mattiev and Kavsek [15], proposed a classification approach based on selecting ‘‘strong’’ class association rules determined by overall learning set coverage. Their method was compared to eight established classification algorithms, including decision tree, rule-based, and Bayesian techniques. And using Adult and Tic-Tac-Toe datasets among others. While achieving a competitive accuracy of 84.9% and the lowest average rule count, the method’s susceptibility to irrelevant or noisy attributes in high-dimensional datasets could lead to overfitting, limiting its applicability to certain data domains.

In Bequé and Lessmann [28] we find an Extreme Learning Machine for credit scoring. In their paper, they investigate ELM’s usability, efficiency, and forecast accuracy systematically and compare them against established scoring techniques. Results suggest that ELM-based credit-scoring models can perform better both in predictive accuracy and resource efficiency. However, ELM is sensitive to the choice of input weights and biases. While an optimal combination brings very good performance for some dataset, it deteriorates under changes in data distribution.

Classification rules were produced with Particle Swarm Optimization (PSO); the integration of SOM and LVQ based neural networks were also used, according to the research of Santana et al. [16]. Despite its potential, this approach can be susceptible to overfitting due to model complexity and high computational costs. Additionally, reliance on PSO might result in suboptimal solutions if not meticulously tuned. Alternatively, Xu et al. [29] proposed a hybrid ensemble credit scoring model based on Generalized Fuzzy Soft Sets (GFSS) theory, integrating feature selection and ensemble learning classification algorithms. Their model demonstrated superior classification accuracy and reliability compared to traditional credit scoring models.

B. QUANTUM-INSPIRED GENETIC ALGORITHM IN LITERATURE

The genetic algorithm is part of the family of evolutionary algorithms. Inspired by the theory of the evolution of species, it takes inspiration from the stages of reproduction at the genetic level to carry out exploratory research work [18], [30]. Its robustness and adaptability [5], [18], [30] motivate us to use it for data-mining.

However, the genetic algorithm in its classic version is constrained by the limitations resulting from the optima problem [31]. In our case, this very quickly becomes debilitating due to the multidimensionality of the datasets studied, which will give a host of different solutions that are not necessarily optimal. Hence the need to regulate convergence towards one solution rather than another.

There are several variants of the genetic algorithm in the literature [30], including the quantum-inspired genetic algorithm [30], [32]. This version allows, among other things, to broaden the field of search for solutions compared to a classical genetic algorithm. The probabilistic nature of the representation of individuals as well as the moderate convergence imposed by the quantum operators allow a more global scanning of the possibilities and therefore a better chance of finding the global optimum [32].

1) GENERAL REPRESENTATION OF THE QUANTUM INDIVIDUAL

In a quantum inspired genetic algorithm the solutions are represented by chromosomes whose atomic unit is no longer the bit as in a classical genetic algorithm but the Q-bit [32]. This representation directly inspired by the principles of quantum physics gives the algorithm its probabilistic property. In fact, where the classic bit can take only one value (0 or 1), the Q-bit can take two values at the same time (0 and 1). In this case it is the superposition of quantum states that we are interested in. These quantum states are represented using bracket notation:

$$\mid \psi\rangle =\mid \alpha\rangle + \mid \beta\rangle \tag{1}$$

where $|0\rangle$ represents 0 in a classical bit and $|1\rangle$ represents 1 in a classical bit, and $\mid \psi\rangle$ denotes more than one vector $\overrightarrow{\psi}$ in a vector space.

The normalization of the states in units guarantees us that:

$$\mid \alpha \mid^2 + \mid \beta \mid^2 = 1 \tag{2}$$

where α and β are two numbers that represent the probability magnitudes of having 0 or 1 respectively.

2) DESCRIPTION OF THE QIGA IN THE LITERATURE

A quantum-inspired genetic algorithm (QIGA) shares the same phases as a classical genetic algorithm which are initialization, selection, evaluation, crossover and mutation [32]. It differs, however, from the latter in the quantum phases which are the measurement, the interference and entanglement [32], [33].

Multiple variants in the construction of the quantum-inspired genetic algorithm can be found in the literature [34], as well as in the choice of the use of the steps mentioned above and of their order. This is relative to the type of problem and the needs of the problem at hand. However, a general architecture emerges that can be modeled as follow in Algorithm named "Quantum-inspired genetic algorithm".

```
Quantum-inspired genetic algorithm
Initialization of the population
repeat
    Selection of best individuals
    Application of genetic operators: Crossover, Mutation
    Application of quantum operators: Interference, Measurement
    Evaluation of the population's fitness
    Saving individual with the best fitness
    Replacement
until stopping criterion is met
return best
```

This passage dives into the various stages of a Quantum-Inspired Genetic Algorithm (QIGA), highlighting its similarities and differences compared to a classical genetic algorithm. Here's a detailed breakdown of each step:

1) Initilization:
   Classical GA: Individuals in the initial population each have a final, fixed value for every attribute. These values are typically chosen randomly within the allowed range for each attribute [18].
   QIGA: While the concept of initialization remains similar, QIGA focuses on the probabilistic nature of quantum mechanics. Instead of fixed values, QIGA initializes the proportions (or weights) associated with each possible quantum state for every attribute in each individual. This essentially represents the likelihood of each attribute taking on a specific value throughout the optimization process [33].
2) Selection: Both QIGA and classical GAs employ a selection mechanism. This phase involves choosing individuals from the current population to participate in further operations like crossover and mutation. The selection process typically prioritizes individuals with higher fitness scores, meaning those that perform better based on the problem being addressed [18], [33].
3) Crossover: Crossover is a fundamental concept in genetic algorithms. It allows the creation of new individuals by combining genetic information (values or probabilities in QIGA's case) from existing parent individuals. The passage acknowledges various crossover techniques exist, but doesn't delve into specific methods applicable to QIGA [18].
4) Mutation:
   Classical GA: Mutation typically involves randomly changing the value of a single attribute within an individual to a different value from the allowed range. This helps introduce diversity and potentially explore new regions of the solution space [18].
   QIGA Mutation (Distinct from Classical GA): Due to the probabilistic nature of QIGA, mutation works differently. Instead of directly changing values, it swaps the probabilities associated with each possible value within an attribute [33]. This maintains the probabilistic representation and allows exploration of different value combinations while keeping the total probability distribution intact.
5) Quantum interference: This stage is a key step that differentiates between QIGA and classical GAs. Quantum interference leverages the principles of quantum mechanics to manipulate the probability distribution of the solutions [32], [35]. Specific quantum operations (represented by "quantum gates") are applied to modify the proportions associated with each possible state for all attributes in the population [32], [35]. The goal is to strategically steer the search towards regions containing potentially better solutions.
6) Measurement: This crucial step marks the transition from the probabilistic values to the discrete values. In QIGA, measurement involves fixing the actual values for each attribute in each individual based on the previously calculated probabilities [35].

This essentially ''collapses'' the probabilistic wave function, determining the final state of the individuals before evaluation [35].

7) Assessment: Similar to classical GAs, QIGA relies on a fitness function to assess the quality of each individual solution. This function is problem-specific and evaluates how well an individual performs in the context of the optimization task at hand [35]. Based on the fitness scores, better individuals are chosen for the next generation [35].
8) Replacement: The final stage involves a selection process where low-performing individuals are replaced with the best ones from the current generation. This ensures that the next iteration builds upon the most promising solutions, gradually improving the overall population towards the optimal solution [18], [32], [35].

3) EVOLUTIONARY ALGORITHM AND QUANTUM INSPIRATION

In Business Intelligence we find the Evolutionary algorithm in several works. We cite for example S. Srinivasan and T. Kamalakannan [36]. Their works focus on the use of the multi-objective genetic algorithm (MOAG) in the prediction and analysis of financial data. The results are very encouraging in the discovery of unique rules as well as the best rules. Bangyal et al. [17] emphasize the importance of balancing exploration (search for new areas of the search space) and exploitation (refining existing solutions) in optimization algorithms. This balance is a strong point of our approach. In our approach, we explore the domain in two phases: the first phase uses the quantum-inspired genetic algorithm (QIGA) for fast attribute selection, while the second phase improves the representation of categorical data to improve the interpretability of the rules. This two-phase strategy allows our approach to be both a robust exploratory method and optimized for solving various types of research problems.

There are many uses of quantum inspired genetic algorithm in various problems. In Dahi et al. [35], the quantum variant gave better results than the classical genetic algorithm for the antenna positioning problem which is a combinatorial optimization problem. In Laboudi and Chikhi [33] a comparison of the two algorithms was made, for this the problem of Knapsack was chosen. The results favor the use of the quantum variant when the number of objects in the Knapsack is large without disadvantaging the latter when the number of objects is small. This motivated us to use the quantum-inspired genetic algorithm for our problem which has similar characteristics.

III. CONTRIBUTION

In this section, we will present our proposal to generate a set of efficient and workable classification rules. We will first share our motivations and what led us to choose this solution for this type of problem. Then, in a second part, we will explain the details of the different steps, the adaptations and the optimizations that our variant of the quantum-inspired genetic algorithm can bring in the context of the creation of classification rules.

A. MOTIVATION

Our first motivation relates to the properties of the quantum-inspired genetic algorithm. The mechanisms of population-based solutions seem to us appropriate for our problem for the following reasons:

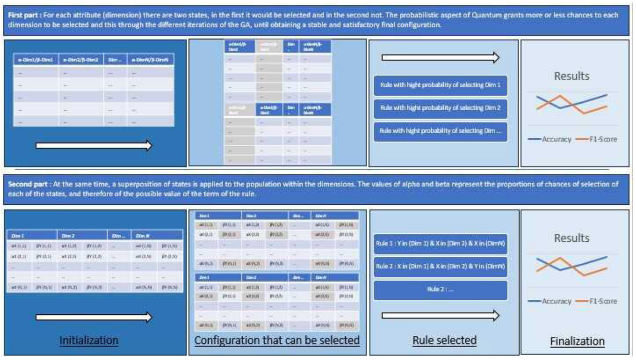
1) Each individual representing a rule, a random initialization of a population allows us to propose several different approaches to tackle the problem. In a context where we have to search in a large dataset, it is interesting to start from several points and be able to combine the best parts of each of our rules via crossover mechanisms to converge towards a good solution.
2) A second property is the robustness of genetic algorithms. In a BI context, we must consider the multitude of data sources that a company seeks to exploit. These data sources not all having good reliability, the need for a robust algorithm is necessary to overcome the problem of biased datasets [5].
3) Finally, the property of acceptance of multiple states of quantum algorithms allows us a wide spectrum of exploration as well as avoid premature convergence of a solution.

Our second motivation relates to the field of classification rules. Classification rules have two properties that are difficult to approach for a quantum-inspired genetic algorithm. Our variant provides a concrete adaptation of the algorithm to the needs of the domain:

1) The first property is that the nature of the values that the fields present in the antecedent of the rule can take. In fact, the QIGA was designed to work on binary values, however in a context of classification rules values can be both binary and categorical (with continuous ones needing to be discretized for optimization purposes).
2) The second property being multi-dimensionality. CRs in many cases have a large number of attributes in the conditional part of the rule. This primarily affects the complexity of the computation and makes optimization difficult and expensive. Even in the case where this is feasible, the result would be rules that are difficult to read for a business actor who wants to use them.

Our variant must therefore offer the possibility of reducing dimensionality, supporting categorical values, this is accomplished by applying the same optimization process which is that of quantum-inspired genetic algorithms.

In the works of Dahi and al. [35] the QIGA tackles the problem of premature convergence by offering a new formula for calculating the interference. This formula is interesting in the sense that even in a large number of iterations, the principle of superposition of states is not compromised and therefore prevents us from falling into local optima.



First part : For each attribute (dimension) there are two states, in the first it would be selected and in the second not. The probabilistic aspect of Quantum grants more or less chances to each dimension to be selected and this through the different iterations of the GA, until obtaining a stable and satisfactory final configuration.

Second part : At the same time, a superposition of states is applied to the population within the dimensions. The values of alpha and beta represent the proportions of chances of selection of each of the states, and therefore of the possible value of the term of the rule.

FIGURE 1. Generation-process.

However, this solution, like those of most works inspired by quantum physics, is suitable for problems whose values are binary.

B. DETAIL OF OUR PROPOSAL

Based on this analysis we propose a solution that implements two processes running at the same time. One to generate a set of rules with a reduced number of attributes. Another to transform an optimization designed for a binary representation of the individual and adapt it to a categorical representation of the latter. We will call the first process Generation-Process, and the second Adaptation-Process.

In Generation-Process we use the QIGA to tackle the high dimensionality and offer a solution that allows us to choose the most relevant attributes and produce readable and understandable rules, in other words a usable knowledge base. This process runs at the heart of Mining and assures that our attribute selection is optimized with the optimization of our rules [7].

The Generation-Process will be divided into two parts: In the first part, we can imagine that for each attribute there are two states, in the first it would be selected and in the second not. If it is selected, we attribute to its state the value of 1 and if it is not, the value of 0. This makes the QIGA applicable on such a sub-problem and we can therefore initiate a population parallel to our population which constitutes the individuals representing our classification rules.

In the second part, we can apply to this parallel population a superposition of the states by values of α and β representing the proportions of chances of selection of each of the states after measurement. Also, we can implement the principle of interference by applying the optimized formula of the work of Dahi [35], as well as all the processes of the classical genetic algorithm in order to obtain the best combination and find the most relevant attributes. This is illustrated in the figure (Fig. 1).

In Adaptation-Process we will adapt the QIGA in order to work on categorical values, to do this we must:

1) For each attribute, the number of values that can be admitted is counted. We convert this number into binary type as described in the work of Freitas [13].
2) Then apply to each bit a superposition of states which will make it a quantum bit. Thus, we can exploit the characteristics of a quantum genetic algorithm such as quantum interference on categorical data.
3) Finally processing, after having measured and fixed our quantum-bit. We will reconstruct our binary numbers, transform them into positive integers and then map each positive integer with the corresponding categorical value according to the position of the latter.

As for the attributes having continuous values, they will be discretized by the entropy method described in the work of Kohavi and Sahami [37].

C. STEPS OF OUR PROPOSED ALGORITHM

During our implementation and for both processes, whether that of optimizing the selection of attributes or that of optimizing values, we will choose the same approach for representing individuals which is the Michigan Approach [7]. The reason is that it is the most natural for a Data-mining task like in our present case.

The application of our variant of QIGA is done by implementing the processes below.

1) INITIALIZATION

We initialize two distinct quantum populations:

The first is the attribute population. In this population each individual represents the terms of the rule, where each term is a superposition of states. Each state describes the probabilities of selecting one of the dimensions of the dataset. The fewer dimensions retained after measurement in the dataset for an individual, the fewer terms will have the rule that we are going to build from this same individual. The selection probabilities of dimensions or attributes must obey the following rule:

$$\mid \alpha \mid^2 + \mid \beta \mid^2 = 1\{\alpha, \beta \in R \tag{3}$$

In this case α represents the probability in which the dimension is selected and β the probability of not selecting the dimension. The figure (Fig. 2) illustrates a typical individual from the first population.

$$\text{Q-I 1} \left\{ \begin{array}{lllll} \alpha_{dim_1} & \alpha_{dim_2} & \alpha_{dim_3} & \cdots & \alpha_{dim_n} \\ \beta_{dim_1} & \beta_{dim_2} & \beta_{dim_3} & \cdots & \beta_{dim_n} \end{array} \right.$$

FIGURE 2. Quantum individual representation.

The second population is that of individuals representing the values that each dimension can include. We will take into account the attributes having categorical values, those having continuous values will be treated in the discretization part. Each rule can have N dimensions, with N depending on the number of non-continuous attributes in the dataset. Each dimension admits M values of categorical or binary type. Each value will be mapped with a numeric identifier corresponding to its position among the M values. The figure (Fig. 3) shows an example of the mapping.

Categorical representation	Numeric representation	Binary representation
Single	1	001
Engaged	2	010
Married	3	011
Widow	4	100

FIGURE 3. **Representation of an example of four values of a dimension N.**

Representation of Values in Bit	Representation of Values in Q-Bit
Value 1 : 001	α_{bit1} α_{bit2} α_{bit3} β_{bit1} β_{bit2} β_{bit3}
Value 2 : 010	
Value 3 : 011	
Value 4 : 100	
Value 5 : 101	

FIGURE 4. **Representation of an example of a dimension of five categorical values.**

An individual represents the N terms with in each term a superposition of the M values specific to each dimension. This superposition is done through the transformation of integer identifiers into binary numbers. Being binary, we can superimpose the probabilities of the Bits that comprise them thus passing from Bit to Quantum-Bit. The figure (Fig. 4) is an example.

The figure (Fig. 5) shows a typical individual of the second population representing the values of the fields of our rules.

The two quantum populations are initiated randomly in order to broaden our field of research and with an equal size in order to be able to combine the two populations in the following processes.

After that, we repeat the following steps until the stopping criterion is met:

2) SELECTION

After initialization of the two quantum populations, we will proceed to the selection by best individuals of the eighty percent of the population which will represent the parents. We select the same parents in both populations in order to be able to combine them at all.

3) CROSSOVER

A simple one-point crossover will be applied to the two quantum populations separately with a seventy percent crossover execution ratio. The resulting individuals will be added to their previously selected parents. The same crossover with the same parents will be carried out in the first and the second population.

4) MUTATION

The mutation in a quantum algorithm is a permutation of the probabilities of α and of β, which we will apply to all the populations with an execution ratio of 1% per iteration. As in crossover, if mutation occurs, we will sum the mutated and unmutated individuals of the two quantum populations separately.

$$\text{Q-I 2} \left\{ \begin{array}{l} \alpha_{bit1-dim1} \cdots \alpha_{bitm1-dim1} \mid \cdots \mid \alpha_{bitmn-dimn} \\ \beta_{bit1-dim1} \cdots \beta_{bitm1-dim1} \mid \cdots \mid \beta_{bitmn-dimn} \end{array} \right.$$

FIGURE 5. **Quantum individual representation with multiple dimensions.**

5) QUANTUM INTERFERENCE

We will apply the optimized quantum interference formula described in the work of Dahi [35] on each Quantum-Bit of the two populations. In the first it will influence the probability of selection of an attribute and in the second it will influence the value that the attribute will take if it is selected.

6) MEASURE

The measurement is the phase during which the Q-bits are transformed into classic Bits. We then combine the two populations by assigning the fixed value of the second population to the corresponding field of the same individual of the first population and whose attribute is selected. In other words, the attribute having a fixed value of "One" in an individual will have its field taken into account in the rule and will have the value measured in the second population.

7) DISCRETIZATION AND EVALUATION

For continuous attributes, we will apply the entropy discretization described in Kohavi and Sahami [37] in order to be able to calculate the fitness of each individual. It should be noted that the discretization step is usually placed higher, except that in our case having quantum values, they must first be fixed in order to be able to use them in the calculation of the discretization by entropy. The objective function used is as followed in equation 4.

$$Q = \frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN} \tag{4}$$

where Q is the measure of Quality.

TP (true positives) is the number of cases covered by the rule that have the class predicted by the rule.

FP (false positives) is the number of cases covered by the rule that have a class different from the class predicted by the rule.

FN (false negatives) is the number of cases that are not covered by the rule but that have the class predicted by the rule.

TN (true negatives) is the number of cases that are not covered by the rule and do not have the class predicted by the rule.

8) REPLACEMENT

After all of the quantum genetic algorithm steps are executed. We will retain the set of quantum individuals of the two populations whose rules have obtained the best fitness score for the next iteration.

9) CONSTRUCTION OF THE RULE-SET

At the end of each iteration, the best rule is retained to build a final set of classification rules.

Finally, we return the best solution found during the optimization process.

IV. RESULTS

A. DATASETS USED

In this section we will introduce the datasets that we utilized for our study in order to assess the effectiveness of our proposed classification models. We have chosen a variety of datasets, from the Machine Learning Repository UCI [40] and The UCI KDD Archive [39], which have been extensively employed in the field of machine learning and data mining. The following datasets were included in our analysis:

1) Adult; The dataset provided here is a collection of data, about people and their income levels. It focuses on determining whether an individual earns than $50,000 or not. The information was gathered by the United States Census Bureau back in 1994 [40].
2) Bank approval; This dataset revolves around individuals and their loan approval status. It aims to classify whether someone was granted a loan or not. The data was collected by a bank within the United States [40].
3) Credit; Here we have a dataset that deals with individuals and their credit scores classifying them as either good or bad. The information comes from a credit bureau located in the United States [40].
4) Australian; This particular dataset provides details, about professions people have in Australia. It focuses on class classification to categorize individuals based on their occupation. The data was sourced from the government [39].
5) German; This dataset is a multi-class classification dataset that contains information about individuals and their credit risk [39].
6) Tic-Tac-Toe; This dataset is a small binary classification dataset that contains information about the outcome of Tic-Tac-Toe games. It was created by the UCI Machine Learning Repository [39].
7) Teacher assistant evaluation; This dataset is a multi-class classification dataset that contains information about USA University students' evaluation [39].
8) Contraceptive method choice; This dataset is a multi-class classification dataset that contains information about women's choices of contraceptive methods. It was collected by the World Health Organization [39].

The datasets were selected because they have been used in previous research. We aim to evaluate the performance and generalizability of our proposed classification models across different domains and problem contexts.

B. EVALUATION METRICS

To assess the performance of our proposed model, we employed several evaluation metrics. These metrics provide insights into different aspects of the model's performance and effectiveness. The following metrics were included in our work:

1) Accuracy; it is a fundamental metric used to assess the overall correctness of a model's predictions. It quantifies the ratio of correctly classified instances to the total number of instances present in the dataset [41]. This metric serves as an indicator of precision and is conventionally reported as a percentage in our findings.
2) Accuracy-rate; it also referred to as the true positive rate (TPR) or recall, gauges the model's proficiency in accurately identifying positive instances. This measure is determined by the ratio of correctly classified positive instances to the total number of actual positive instances within the dataset [11]. Similar to accuracy, this metric signifies precision and is expressed as a percentage in our reported outcomes.
3) F-score; this composite measure is calculated as the harmonic mean of precision and recall. Precision denotes the fraction of correctly classified positive instances, while recall represents the fraction of actual positive instances correctly classified by the model. Particularly advantageous when dealing with imbalanced datasets, the F-score comprehensively accounts for false positives and false negatives [41]. The F-score exemplifies the model's reliability and is presented as a ratio between 0 and 1 in our reported results.
4) Kappa; or Cohen's kappa coefficient, scrutinizes the level of agreement between a model's predictions and the ground truth, surpassing what would be anticipated by chance alone. This coefficient is computed as the disparity between the observed agreement and the agreement expected by chance, normalized by the maximum feasible agreement. Kappa values in proximity to 1 signify a heightened degree of concordance between the model's predictions and the ground truth [41]. The Kappa coefficient stands as an indicator of reliability and is reported as a ratio between 0 and 1 in our findings.

C. DESCRIPTION OF RESULTS

To validate the efficacy of our proposed approach, a series of experiments followed the analysis on a collection of benchmark datasets widely used in state-of-art research. The performance of the obtained models was evaluated in terms of a set of commonly used criteria such as, Accuracy (Acc), Accuracy Rate (Acc-R), F-score and Kappa.

Table 1 provides the comparison of our method with existing techniques on the Credit Approval dataset. We compare the findings of our study with Archer et al. [42], Seidlova et al. [11], and Weng et al. [19]. The same is again true for Table 2 and Table 3 presenting comparative results for the Bank Marketing and Adult datasets respectively while including further baselines by Guo et al. [20], Tripathi et al. [14], and Mattiev et al. [15].

TABLE 1. **Table that exposes our solution in comparison with others in the financial field for Credit Approval dataset.**

	Acc	F-score	kappa	Acc-R	References
Zero-Rule	/	/	/	55,5	J Archer et al. [42]
Random Forest	/	/	/	83,9	J Archer et al. [42]
C4.5	/	/	/	87,2	J Archer et al. [42]
ACO	/	/	/	89,9	R Seidlova et al. [11]
CBDT	94,00	0,93	/	/	C Weng et al. [19]
CBMLP	80,00	0,78	/	/	C Weng et al. [19]
CBNB	69,00	0,64	/	/	C Weng et al. [19]
CBSVM	84,00	0,67	/	/	C Weng et al. [19]
QIGA 1st	93,84	0,90	0,82	93,48	/
QIGA 2nd	90,80	0,96	0,85	91,50	/

Number of attributes: 15. Number of instances: 690.

TABLE 2. **Table that exposes our solution in comparison with others in the financial field for Bank Marketing dataset.**

	Acc	F-score	kappa	Acc-R	References
Zero-Rule	/	/	/	88,3	J Archer et al. [42]
Random Forest	/	/	/	89,7	J Archer et al. [42]
C4.5	/	/	/	90,3	J Archer et al. [42]
ACO	/	/	/	91,9	R Seidlova et al. [11]
EFO-ANN	88,76	/	/	/	W Guo et al. [20]
FSA-ANN	88,76	/	/	/	W Guo et al. [20]
HSA-ANN	88,22	/	/	/	W Guo et al. [20]
SSD-ANN	88,76	/	/	/	W Guo et al. [20]
TDNN-ILSF	96,99	/	/	/	D Tripathi et al. [14]
QIGA 1st	91,15	0,04	0,04	15,61	/
QIGA 2nd	82,74	0,90	0,25	71,82	/

Number of attributes: 16. Number of instances: 45212.

TABLE 3. **Table that exposes our solution in comparison with others in the financial field for Adult dataset.**

	Acc	F-score	kappa	Acc-R	References
Zero-Rule	/	/	/	75,90	J Archer et al. [42]
Random Forest	/	/	/	84,30	J Archer et al. [42]
C4.5	/	/	/	86,20	J Archer et al. [42]
ACO	/	/	/	87,90	R Seidlova et al. [11]
DTNB	73,00	/	/	/	J Mattiev et al. [15]
DT	82,00	/	/	/	J Mattiev et al. [15]
PT	82,10	/	/	/	J Mattiev et al. [15]
FR	75,20	/	/	/	J Mattiev et al. [15]
RDR	80,80	/	/	/	J Mattiev et al. [15]
J&B	80,80	/	/	/	J Mattiev et al. [15]
QIGA 1st	85,25	0,11	0,60	43,6	/
QIGA 2nd	73,22	0,63	0,60	71,1	/

Number of attributes: 14. Number of instances: 32561.

TABLE 4. **Table of parameters for QIGA 1st Line of financial datasets.**

Parametre	Value
Population size	500
Selection pourcentage	80
Crossover rate	70
Mutation rate	1
LR	0.1
Number of generation	300
Minimum number of case per rule	5
Number of rule user to test	10

TABLE 5. **Table of parameters for QIGA 2nd Line of financial datasets.**

Parametre	Value
Population size	800
Selection pourcentage	1
Crossover rate	70
Mutation rate	1
LR	0.1
Number of generation	1000
Minimum number of case per rule	5
Number of rule user to test	10

TABLE 6. **Table that exposes our solution in comparison with others in context of binary class dataset for the Australian Credit Approval dataset.**

	Acc	F-score	kappa	Acc-R	References
AntMiner+	84,05	/	/	/	D Martens et al. [38]
AntMiner	84,09	/	/	/	D Martens et al. [38]
AntMiner2	84,3	/	/	/	D Martens et al. [38]
AntMiner3	83,61	/	/	/	D Martens et al. [38]
RIPPER	84,52	/	/	/	D Martens et al. [38]
C4,5	84,82	/	/	/	J Archer et al. [42]
1NN	80,83	/	/	/	D Martens et al. [38]
Logit	84,83	/	/	/	V Vapnik et al. [43]
SVM	85,22	/	/	/	V Vapnik et al. [43]
CBDT	93,00	0,92	/	/	C Weng et al. [19]
CBMLP	78,00	0,78	/	/	C Weng et al. [19]
CBNB	78,00	0,71	/	/	C Weng et al. [19]
CBSVM	94,00	0,94	/	/	C Weng et al. [19]
TDNN-UDFS	92,07	/	/	/	D Tripathi et al. [14]
SVM-R	86,10	/	/	/	A Bequé et al. [28]
LVQ+PSO	86,41	/	/	/	Santana et al. [16]
GFSS	88,10	/	/	/	Xu et al. [29]
QIGA	80,87	0,77	0,54	77,39	/

Number of attributes : 14. Number of instances : 690.

Our proposed approach was evaluated under two distinct parameterizations: the first one optimized for a maximum accuracy (QIGA 1st) and the second one prioritizing a balance between accuracy and reliability (QIGA 2nd). Both of the above results are summarized in Table 1, 2 and 3.

For the purpose of comparison, the number of instances and attributes of each of the datasets is stated in the following tables' footnotes. The detail of the setting of the lengths of tweets and the number of features used is presented in Table 4 and Table 5 for the accuracy optimization and the desired accuracy-reliability pair, respectively.

The same comparison applied to the Australian Credit Approval, German Credit Data and Tic-Tac-Toe datasets is presented in Table 6, 7, and 8. We add new standards by Martens et al. [38], Bequé et al. [28], Santana et al. [16], and Xu et al. [29]. These datasets have the same format, specifically they both use a binary class.

In contrast, Table 9 presents results for the multi-class datasets Contraceptive Method Choice and Teacher Assistant Evaluation. For these datasets, we focus on the comparison with the baseline models from Martens et al. [38].

It should be mentioned that the generalized characteristics of the datasets in tables 6, 7, 8, 9 in comparison with the tables 1, 2 and 3 are the size of the matrix and the parameters; thus, the parameterization used for these datasets was equal to (QIGA 1st) according to the table 4.

In our comparative analysis with the work of David Martens et al. [38], we present supplementary findings in table 10 that shed light on the execution time and average number of terms per generated rule. It is important to note that this comparison is conducted specifically with the work of David Martens et al. [38], and our average time and term

TABLE 7. Table that exposes our solution in comparison with others in context of binary class dataset for German Credit Data dataset.

	Acc	F-score	kappa	Acc-R	References
AntMiner+	71,88	/	/	/	D Martens et al. [38]
AntMiner	71,08	/	/	/	D Martens et al. [38]
AntMiner2	70,66	/	/	/	D Martens et al. [38]
AntMiner3	71,29	/	/	/	D Martens et al. [38]
RIPPER	73,18	/	/	/	D Martens et al. [38]
C4,5	74,2	/	/	/	J Archer et al. [42]
1NN	74,05	/	/	/	D Martens et al. [38]
Logit	75,24	/	/	/	V Vapnik et al. [43]
SVM	73,68	/	/	/	V Vapnik et al. [43]
CBDT	89,00	0,92	/	/	C Weng et al. [19]
CBMLP	73,00	0,84	/	/	C Weng et al. [19]
CBNB	50,00	0,53	/	/	C Weng et al. [19]
CBSVM	86,00	0,91	/	/	C Weng et al. [19]
TDNN-UDFS	87,37	/	/	/	D Tripathi et al. [14]
SVM-R	75,90	/	/	/	A Bequé et al. [28]
LVQ+ varPSO	69,81	/	/	/	Santana et al. [16]
GFSS	87,60	/	/	/	Xu et al. [29]
QIGA	75,14	0,75	0,51	76,04	/

Number of attributes : 19. Number of instances : 1000.

TABLE 8. Table that exposes our solution in comparison with others in context of binary class dataset for Tic-Tac-Toe dataset.

	Acc	F-score	kappa	Acc-R	References
AntMiner+	99,57	/	/	/	D Martens et al. [38]
AntMiner	75,03	/	/	/	D Martens et al. [38]
AntMiner2	71,13	/	/	/	D Martens et al. [38]
AntMiner3	68,94	/	/	/	D Martens et al. [38]
RIPPER	97,99	/	/	/	D Martens et al. [38]
C4,5	83,79	/	/	/	J Archer et al. [42]
1NN	98,5	/	/	/	D Martens et al. [38]
Logit	65,57	/	/	/	V Vapnik et al. [43]
SVM	91,06	/	/	/	V Vapnik et al. [43]
DTNB	69,90	/	/	/	J Mattiev et al. [15]
DT	74,40	/	/	/	J Mattiev et al. [15]
PT	85,20	/	/	/	J Mattiev et al. [15]
FR	94,30	/	/	/	J Mattiev et al. [15]
RDR	94,10	/	/	/	J Mattiev et al. [15]
J&B	95,80	/	/	/	J Mattiev et al. [15]
QIGA	38,75	0,1	0,05	10,05	/

Number of attributes : 9. Number of instances : 958.

per rule values encompass all the datasets used in our study, providing a comprehensive overview of our algorithm's performance in comparison to their work.

V. DISCUSSION

In this section we discuss the different results obtained and presented in the previous section. Indeed, we can advance several observation:

The first one is the number of metrics we used to validate our approach. There are four of them, two of accuracy and which are the accuracy and the accuracy rate. The two others are reliability metrics: Kappa and F-score. This highlights the axes on which we have based our work.

In Table 1, 2 and 3 we have exposed two possible configurations of our algorithm. The first one called (QIGA 1st) is a precision-oriented configuration. With a population size of 500 and a number of 300 generations, we have an interesting combination to perform in precision. This can be seen on the two datasets Bank-Marketing and Adult, where the accuracy achieved is 91.15 for the first one in Table 2 and 85.25 for the second one in Table 3. This puts our algorithm in a good position to look for more performance. However, taking into account the reliability metric, we notice, as expected, that the algorithm strongly loses credibility. This can be seen on the two values of F-score 0.04 and 0.11 on Bank-Marketing in Table 2 and Adult in Table 3 respectively, and on Kappa with values of 0.04 and 0.06 on Bank-Marketing as we can see Table 2 and Adult on Table 3 respectively. The second configuration called (QIGA 2nd) is oriented to perform in precision while keeping a high reliability. With a bigger population size of 800 instances and a number of generations equal to 1000. The algorithm goes less far in its optimization of the accuracy. The accuracy is 82.74 and the accuracy-rate is 71.82 for the dataset Bank-Marketing. And 73.22 for the accuracy and 71.1 for the accuracy-rate of the dataset Adult. However, for both datasets we obtain an excellent reliability of 0.90 and 0.63 of F-score on the first and second dataset respectively.

On the Table 1, the Credit dataset obtains practically the same results on both configurations. A precision of 93.84 and an accuracy-rate of 93.48 with an excellent reliability value of 0.9 in F-score and 0.82 in Kappa for (QIGA 1st). A precision of 90.80 and an accuracy rate of 91.50 and also an excellent reliability value of 0.96 in F-score and 0.85 in Kappa for (QIGA 2nd). These results place our algorithm in first position in comparisons made with the work of R Seidlova et al. [11] when we compare on the Accuracy-Rate, and the second position after CBDT of C Weng et al. [19] when referring both Accuracy or F-Score.

On Table 6 and 7 we report the work of D Martens et al. [38], Vapnik et al. [43], Weng et al. [19], Tripathi et al. [14], Bequé et al. [28], Santana et al. [16] and Xu et al. [29] on the dataset with binary classes. In Table 6, when using the Australian Credit Approval dataset, we obtain a score of 80.87 in accuracy and 0.77 on F-score that makes our QIGA in theard position after both CBDT and CBSVM considering the tuple (accuracy/F-score), and low average position considering the accuracy as the unique performance evaluation key. Conversely, on the German dataset Table 7, our method achieves a competitive accuracy of 75.14, ranking second closely behind the Logit model's 75.24 accuracy when considering direct approaches. Notably, our solution exhibits a robust F-score of 0.75, indicating strong reliability. While demonstrating competitive performance in hybrid approaches across datasets when we are on the hight average position. Our algorithm appears less suited to the Tic-Tac-Toe dataset as we can see on Table 8.

The results obtained on multi-class datasets are described in Table 9. We have the best result in accuracy on the Teacher assistant evaluation dataset, with a score of 58.40 of accuracy in front of the second best which is the AntMiner+ with 56.73 and a reliability 0.75 on the F-score metric.

Regarding the dataset Contraceptive method choice, our solution is the second best and very close to the best which is the SVM. Our accuracy score reaches 48.09 when the SVM is at 48.55. However, we could not obtain an acceptable reliability, because the scores of 0.11 in F-score and 0.16 in Kappa are considered as low.

TABLE 9. Table that exposes our solution in comparison with others in context of multi-class dataset.

	Contraceptive Method Choice[1]				Teacher Assistant Evaluation[2]				
	Acc	F-score	kappa	Acc-R	Acc	F-score	kappa	Acc-R	References
AntMiner+	45,93	/	/	/	56,73	/	/	/	D Martens et al. [38]
AntMiner	42,32	/	/	/	40,39	/	/	/	D Martens et al. [38]
AntMiner2	41,49	/	/	/	43,73	/	/	/	D Martens et al. [38]
AntMiner3	40,85	/	/	/	40,39	/	/	/	D Martens et al. [38]
RIPPER	48,94	/	/	/	35,2	/	/	/	D Martens et al. [38]
C4,5	46,6	/	/	/	47,2	/	/	/	J Archer et al. [42]
1NN	42,16	/	/	/	50,2	/	/	/	D Martens et al. [38]
Logit	47,52	/	/	/	51,96	/	/	/	V Vapnik et al. [43]
SVM	48,55	/	/	/	51,96	/	/	/	V Vapnik et al. [43]
QIGA	48,09	0,11	0,16	44,16	58,40	0,75	0,34	39,77	/

[a] Number of attributes : 9. Number of instances : 1473.

[b] Number of attributes : 6. Number of instances : 151.

TABLE 10. Table that exposes average execution time accompanied by the average term per rule ratio.

	Time	T/R
AntMiner+	1776	3,3
AntMiner	390	1,3
AntMiner2	422	1,3
AntMiner3	867	1,3
RIPPER	0,3	2,8
C4,5	0,1	/
1NN	0,0	/
Logit	0,3	/
SVM	941	/
QIGA	415	5

In addition to the aforementioned analysis, we present supplementary findings in table 10 that provide insights into the execution time and average number of terms per generated rule in comparison with other solutions. Notably, our algorithm maintains a fixed minimum value of 5 terms, which ensures a consistent baseline for rule generation. Consequently, our algorithm consistently strives to optimize other performance metrics while adhering to this predetermined term count. This attribute distinguishes our solution as the superior method in terms of generating terms per rule. Furthermore, the recorded average execution time of our algorithm is 415 seconds, positioning it as the second most time-efficient approach among the algorithms for which the number of terms per rule generated is known. This information holds significance as it contributes to the credibility and reliability of our results. Additionally, our algorithm secures the 6th position in the overall ranking based on execution time, underscoring its competitive performance in this aspect.

Furthermore, our quantum-inspired approach demonstrates its relevance by outperforming both the AntMiner+ algorithm and the basic AntMiner in terms of generating a higher number of terms per rules. Specifically, our approach achieves a 1.5 times increase compared to AntMiner+ and a remarkable 3.8 times improvement compared to AntMiner as can be seen in table 10. Remarkably, despite these significant gains in terms of term generation, our approach maintains superior performance to AntMiner+ in three out of the five evaluation tests.

The success of our quantum-inspired approach can be attributed to the implementation of a quantum-inspired genetic algorithm. A key factor in our algorithm's effectiveness lies in its unique probabilistic representation, which operates not only at the level of values but also, crucially, at the level of dimensions. This allows us to make decisions regarding the selection or exclusion of dimensions during each iteration. Importantly, our algorithm assigns a high probability to selecting the most favorable choice, while still allowing for the possibility of retaining the alternative choice. This probabilistic aspect is challenging to implement in alternative approaches, as their convergence tends to be stronger, making it difficult to eliminate dimensions entirely.

Finally, in our algorithm, the population size and number of generations serve as influential factors. Increasing the population size and the number of generations generally leads to improved accuracy of our algorithm. However,

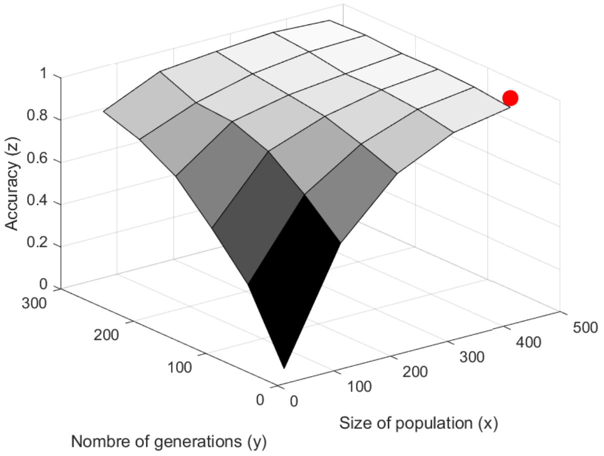


FIGURE 6. Influence of the number generation and population size on accuracy.

it is important to note that beyond certain threshold values, specifically 200 for the number of generations and 400 for the population size, the convergence becomes slower, resulting in excessively long calculation times. Moreover, Table 1, 2 and 3 highlights the observed trade-off between reliability and accuracy, as increasing these factors does not necessarily yield a significant improvement in accuracy while compromising reliability.

To visually depict the relationship between these factors and our algorithm's performance, Figure (Fig. 6) illustrates the progression of accuracy with respect to the population size and number of generations. Meanwhile, Figure (Fig. 7) showcases the impact of these factors on computation time. These graphical representations provide a clear visualization of the trends and help in identifying the optimal values for our algorithm.

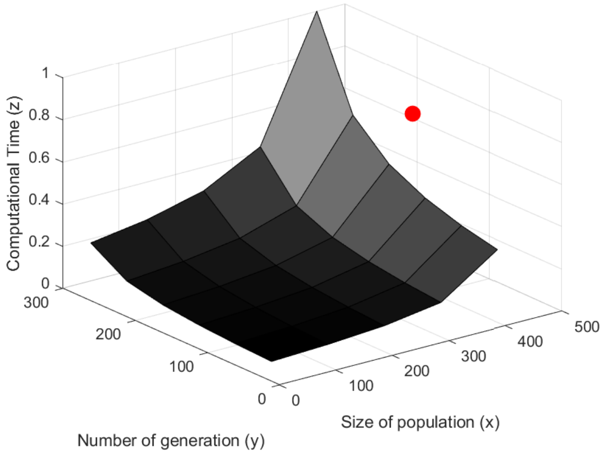


FIGURE 7. **Influence of the number generation and population size on execution time.**

VI. CONCLUSION

This paper introduces a novel variant of the quantum genetic algorithm specifically tailored for classification rules within the context of Business Intelligence (BI). Our variant exploits the power of a proven optimization in a binary problem field, by adapting and propagating it successfully to tackle the dimensionality problem in a categorical problem field.

A key contribution of this work resides in the algorithm's ability to simultaneously optimize both precision and reliability. By carefully configuring the QIGA, we demonstrate the possibility to achieve high accuracy while maintaining strong model reliability, a critical aspect often neglected in other studies. Our empirical evaluation on multiple datasets reveals that the proposed QIGA challenges the best existing methods in terms of accuracy rate, ranking second after CBDT in terms of both precision and F-score. Moreover, the algorithm's ability to handle high-dimensional data effectively contributes to the generation of concise and interpretable rule sets.

As it has been illustrated in this paper, the proposed QIGA has a favorable performance, but the computational time is an issue that hinders the methodology when applied to large-scale datasets. Future studies should be directed towards increasing the effectiveness of the algorithm and at the same time aiming at improving the performance of the algorithm. Therefore, other research directions might include investigating how to improve the results of the QIGA method when it is used in combination with other methods, for example, feature selection or ensemble methods. Lastly, examining the possibility to extend the usage of the proposed method to other domains outside Business Intelligence is an interesting idea for future research.

ACKNOWLEDGMENT

Youcef Islem Adnane would like to express his sincere gratitude to Pr. Amer Draa for his valuable guidance and insights in the field of genetic algorithms. Youcef Islem Adnane would also like to extend his gratitude towards his colleagues: Mohammed Imed Eddine Khaldi and Wayoud Bouzeraib for their invaluable technical advice and input throughout this research endeavor. Their support was instrumental in the successful completion of this work.

REFERENCES

[1] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011.

[2] E. P. Lim, H. Chen, and G. Chen, "Business intelligence and analytics: Research directions," *ACM Trans. Manag. Inf. Syst.*, vol. 3, no. 4, 2013, pp. 1–10.

[3] V. B. Vukšić, M. P. Bach, and A. Popovič, "Supporting performance management with business process management and business intelligence: A case analysis of integration and orchestration," *Int. J. Inf. Manage.*, vol. 33, no. 4, pp. 613–619, Aug. 2013.

[4] G. R. Gangadharan and S. N. Swami, "Business intelligence systems: Design and implementation strategies," in *Proc. 26th Int. Conf. Inf. Technol. Interfaces*, Jun. 2004, pp. 139–144.

[5] N. Dedić and C. Stanier, "Towards differentiating business intelligence, big data, data analytics and knowledge discovery," in *Innovations in Enterprise Information Systems Management and Engineering* (Lecture Notes in Business Information Processing). London, U.K.: Springer, 2017, pp. 114–122.

[6] T. S. Lee and I. F. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines," *Expert Syst. Appl.*, Vol. 28, no. 4, pp. 743–752, 2005, doi: 10.1016/j.eswa.2004.12.031.

[7] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[8] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–7.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[10] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, "Classification-rule discovery with an ant colony algorithm," in *Encyclopedia of Information Science and Technology*, 1st ed. Hershey, PA, USA: IGI Global, 2005, pp. 420–424.

[11] R. Seidlova, J. Poživil, and J. Seidl, "Marketing and business intelligence with help of ant colony algorithm," *J. Strategic Marketing*, vol. 27, no. 5, pp. 451–463, Jul. 2019.

[12] S. Dehuri, S. Patnaik, A. Ghosh, and R. Mall, "Application of elitist multi-objective genetic algorithm for classification rule generation," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 477–487, Jan. 2008.

[13] A. A. Freitas, *Data Mining and Knowledge Discovery With Evolutionary Algorithms*. Springer, 2002.

[14] D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla, and B. R. Reddy, "Experimental analysis of machine learning methods for credit score classification," *Prog. Artif. Intell.*, vol. 10, no. 3, pp. 217–243, Sep. 2021.

[15] J. Mattiev and B. Kavsek, "Coverage-based classification using association rule mining," *Appl. Sci.*, vol. 10, no. 20, p. 7013, Oct. 2020.

[16] P. J. Santana, A. V. Monte, E. Rucci, L. C. Lanzarini, and A. F. Bariviera, "Analysis of methods for generating classification rules applicable to credit risk," Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina, Tech. Rep., 2017.

[17] W. H. Bangyal, R. Shakir, N. U. Rehman, A. Ashraf, and J. Ahmad, "An improved seagull algorithm for numerical optimization problem," in *Proc. Int. Conf. Swarm Intell.* Shenzhen, China: Springer, 2023, pp. 297–308.

[18] A. Anwaar, A. Ashraf, W. H. K. Bangyal, and M. Iqbal, "Genetic algorithms: Brief review on genetic algorithms for global optimization problems," in *Proc. Hum.-Centered Cogn. Syst. (HCCS)*, Dec. 2022, pp. 1–6.

[19] C.-H. Weng and C.-K. Huang, "A hybrid machine learning model for credit approval," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 1439–1465, Dec. 2021.

[20] W. Guo, Y. Yao, L. Liu, and T. Shen, "A novel ensemble approach for estimating the competency of bank telemarketing," *Sci. Rep.*, vol. 13, no. 1, p. 20819, Nov. 2023.

[21] S. M. Monzurur Rahman, M. Rokibul Alam Kotwal, and X. Yu, "Mining classification rules via an apriori approach," in *Proc. 13th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2010, pp. 388–393.

[22] A. Satheesh, D. K. Mishra, and R. Patel, "Classification rule mining for object oriented databases: A brief review," in *Proc. 1st Int. Conf. Comput. Intell., Commun. Syst. Netw.*, Jul. 2009, pp. 259–263.

[23] A. A. Freitas, "Understanding the crucial differences between classification and discovery of association rules: A position paper," *ACM SIGKDD Explor. Newslett.*, vol. 2, no. 1, pp. 65–69, Jun. 2000.

[24] Y. Chen, L. Chen, and L. Tu, "Parallel ant colony algorithm for mining classification rules," in *Proc. IEEE Int. Conf. Granular Comput.*, May 2006, pp. 85–90.

[25] E. Noda, A. A. Freitas, and H. S. Lopes, "Discovering interesting prediction rules with a genetic algorithm," in *Proc. Congr. Evol. Comput. (CEC)*, vol. 2, Jul. 1999, pp. 1322–1329.

[26] D. J. Hand, "Assessing the performance of classification methods," *Int. Stat. Rev.*, vol. 80, no. 3, pp. 400–414, Dec. 2012.

[27] A. A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," in *Advances in Evolutionary Computing: Theory and Applications*. Springer, 2003, pp. 819–845.

[28] A. Bequé and S. Lessmann, "Extreme learning machines for credit scoring: An empirical evaluation," *Expert Syst. Appl.*, vol. 86, pp. 42–53, Nov. 2017.

[29] D. Xu, X. Zhang, and H. Feng, "Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model," *Int. J. Finance Econ.*, vol. 24, no. 2, pp. 903–921, Apr. 2019.

[30] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimedia Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021.

[31] S. H. Aljahdali, A. S. Ghiduk, and M. El-Telbany, "The limitations of genetic algorithms in software testing," in *Proc. ACS/IEEE Int. Conf. Comput. Syst. Appl. - AICCSA*, May 2010, pp. 1–7.

[32] R. Lahoz-Beltra, "Quantum genetic algorithms for computer scientists," *Computers*, vol. 5, no. 4, p. 24, Oct. 2016.

[33] Z. Laboudi and S. Chikhi, "Comparison of genetic algorithm and quantum genetic algorithm," *Int. Arab J. Inf. Technol.*, vol. 9, no. 3, pp. 243–249, May 2012.

[34] G. Zhang, "Quantum-inspired evolutionary algorithms: A survey and empirical study," *J. Heuristics*, vol. 17, no. 3, pp. 303–351, Jun. 2011.

[35] Z. A. E. M. Dahi, C. Mezioud, and A. Draa, "A quantum-inspired genetic algorithm for solving the antenna positioning problem," *Swarm Evol. Comput.*, vol. 31, pp. 24–63, Dec. 2016.

[36] S. Srinivasan and T. Kamalakannan, "Multi criteria decision making in financial risk management with a multi-objective genetic algorithm," *Comput. Econ.*, vol. 52, no. 2, pp. 443–457, Aug. 2018.

[37] R. Kohavi and M. Sahami, "Error-based and entropy-based discretization of continuous features," in *Proc. KDD*, 1996, pp. 114–119.

[38] D. Martens, M. De Backer, R. Haesen, J. Vanthienen, M. Snoeck, and B. Baesens, "Classification with ant colony optimization," *IEEE Trans. Evol. Comput.*, Vol. 11, no. 5, pp. 651–665, Oct. 2007.

[39] S. Hettich and S. D. Bay, "The UCI KDD archive," Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 1996. [Online]. Available: http://kdd.ics.uci.edu

[40] M. Lichman, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Los Angeles, CA, USA, 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[41] Ž. Vujović, "Classification model evaluation metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021.

[42] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Statist. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, Jan. 2008.

[43] V. Vapnik, "The nature of statistical learning theory," *JCS&T*, vol. 17, no. 1, pp. 20–28, 2013.

YOUCEF ISLEM ADNANE was born in Oum El Bouaghi, Algeria, in May 1994. He received the master's degree in NICT from the University of Constantine 2 Abdelhamid Mehri, in 2019, showcasing their dedication and aptitude for the subject.

He is currently an accomplished individual with a passion for the field of new information and communication technology (NICT). In 2019, they secured an impressive second-place ranking in their class by excelling in the national admission competition for doctoral training. Throughout his career, he has gained valuable professional experience through various positions. Currently, based in Algiers, Algeria, he holds the role of a Data Consultant with KPMG, where he leads an IT team in providing strategic solutions for international clients' business intelligence challenges. Simultaneously, for the past four years, he has been an integral part of the laboratory team with MISC, University of Constantine 2 Abdelhamid Mehri. His research interests include within the team has centered around the implementation and utilization of evolutionary approaches in the field of optimization, leveraging data-driven insights, and innovative techniques. He is dedicated to pushing the boundaries of NICT and contributing to advancements in the field.

MOUNIRA ZERARI was born in Algiers, Algeria, in August 1979. She received the bachelor's degree in computer science from the University of Mentouri, Constantine, Algeria, laying the foundation for her future achievements and the Ph.D. degree in computer science from Mentouri University, in 2012, further solidifying her expertise and research capabilities.

Throughout his career, she has amassed a wealth of work experience in academia and research. From 2005 to 2012, he was an Assistant Professor with the Department of Computer Science, Mentouri University, Constantine, Algeria, where he actively contributed to the advancement of knowledge in the field. Then, she took on the role of an Assistant Professor with Abdelhamid Mehri University, Constantine, from 2012 to 2016, further solidifying his expertise and mentoring students, where she holds the esteemed position of an Associate Professor, making significant contributions to the academic community. In addition to his teaching responsibilities, she has published research papers in renowned conferences and journals. Notably, his publications include *Context-Aware Process Mining Framework for Business Process Flexibility* (ACM iiWAS 2010) and *Agent Interaction Protocols in Support of Cloud Services Composition* (Industrial Applications of Holonic and Multi-Agent Systems—Sixth International Conference, HoloMAS 2013). Her research interests include context-aware process mining, business process flexibility, and agent interaction protocols in cloud services composition.

• • •