Gravitational Search Algorithms in Data Mining: A Survey

Poonam1, Saroj Ratnoo2

Research Scholar, Department of Computer Science and Engineering, GJU S&T, Hisar, India1
Professor, Department of Computer Science and Engineering, GJU S&T, Hisar, India2

Abstract: A Gravitational Search Algorithm is a meta-heuristic search algorithm inspired by Newton’s laws of motion and gravitation force. Gravitational search based algorithms have attracted the attention of data mining community due to their ability to effectively search large and complex search spaces as is the case with the domain of data mining and knowledge discovery. Gravitational search algorithms have been applied to the data mining tasks like classification, clustering objects and mining association rules etc. This paper presents a survey on the applications of Gravitational Search based algorithms in data mining and indicates the scope of further research on using GSAs in the field of data mining.

Keywords: Data Mining, Gravitational Search Algorithm (GSA), Classification, Clustering, Association Rule Mining, Feature Selection.

I. INTRODUCTION

A huge amount of data is being collected and stored in databases, data warehouses or other information repositories nowadays. The data is increasing exponentially and it is not humanly possible to analyze all this data to extract some interesting knowledge from it. This has led to the emergence of a field called data mining. Data mining is the automated or semi-automated non-trivial process of extracting interesting information from a large amount of data [1]. Data mining is also referred as Knowledge Discovery in Databases (KDD) [1][2]. The information extracted through data mining algorithms can help in decision making processes. Based on the type of information extracted, data mining can be categorized into descriptive and predictive tasks. A descriptive data mining task deals with general properties of data in a database whereas a predictive data mining task performs inference on the current data in order to make predictions [2]. Descriptive data mining tasks include association rule mining and clustering and predictive data mining tasks include classification and regression. Many nature inspired algorithms like Particle Swarm Optimization (PSO)[3], Genetic Algorithms (GAs)[3], Ant Colony Optimization (ACO)[3], and Gravitation Search Algorithms (GSAs) have been applied to various data mining tasks and have shown better performance than the traditional methods. A GSA, one of the relatively new techniques inspired by Newtonian principles of motion and gravitation force, has shown promising results in the domain of data mining. The main objective of this paper is to provide an up-to-date survey of the applications of the gravitational search algorithms in the field of data mining.

The remainder of the paper is organized as follows: Section 2 gives an overview of various data mining tasks. Gravitational Search Algorithm and its variants are discussed in Section 3. Section 4 presents the application of GSA in classification. The applications of GSA in clustering are elaborated in Section 5. Section 6 describes the applications of GSA in association rule mining. The applications of GSA in feature selection are discussed in Section 7. Section 8 concludes the paper.

II. DATA MINING TASKS

Data Mining can be categorized into many tasks such as classification, feature selection, association rule mining and clustering. A very brief description of each task addressed in this paper is given below.

A. Classification
Classification is a data mining task that deals with building a classification model by analyzing the training data consisting of a set of attributes and the predefined class labels. Later, the classification model is used to predict the class labels of unseen data [2].

B. Clustering
Clustering is a process that partitions the data objects into groups so that the objects within one group are similar to each other and dissimilar to the objects of the other groups [2].
C. Association rule mining
Association rule mining is a data mining task that deals with finding the strong associations and correlations between various data items, most often in transactional databases [2].

D. Feature Selection:
Feature Selection is a data mining task that deals with identifying and selecting a subset of relevant and non-redundant features for the underlying data mining task. Feature selection enhances the accuracy and comprehensibility of the knowledge discovered.

III. GRAVITATIONAL SEARCH ALGORITHM

A Gravitational search algorithm is an optimization algorithm that draws inspiration from Newton’s law of motion and gravity. In this algorithm, search space is initialized with a number of agents called objects. Each object has mass that represent solution for the underlying problem. All objects attract the other objects with a gravitational force. This gravitational force produce movement of low masses objects towards the objects with the heavier masses [4]. In this algorithm, each object has four specifications associated with it, i.e. position, active gravitational mass, passive gravitational mass and inertial mass. The algorithm works by adjusting the gravitational mass and inertial mass of objects are evaluated using a fitness function. [4].

A. Implementing a GSA
The main steps to implement a GSA are given below.

1) Initializing agents:
The positions of the N number of agents are initialized randomly as:

\[ X_i = (x_{i1}, \ldots, x_{id}, \ldots, x_{in}) \text{ for } i=1, 2 \ldots N. \]

Where \( x_{id} \) is the position of the \( i^{th} \) agent in the \( d^{th} \) dimension.

2) Evaluating fitness:
Fitness is evaluated based on the objective function. Best and worst solution depends on the type of problems. For the maximization problems, a solution with the maximum fitness is considered the best solution and vice versa.

3) Computing Gravitational constant (G):
The value of the gravitational constant depends on initial value \( G_0 \) and iteration \( t \). It is calculated as given below.

\[ G(t) = G_0 e^{-\alpha t} \]

Here, the value of \( G_0 \) and \( \alpha \) will be diminished with time to control the search space.

4) Computing Mass:
Mass of each agent is calculated as follows.

\[ m_i(t) = \frac{\text{fit}_i(t) - \text{worst}(t)}{\text{best}(t) - \text{worst}(t)} \]

And

\[ M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \]

Here \( \text{fit}_i(t) \) represents the fitness value of the \( i^{th} \) agent at time\( t \) and \( \text{worst}(t) \) and \( \text{best}(t) \) represent the worst and best fitness at time\( t \).

5) Computing Acceleration

\[ a_i = \frac{F_i(t)}{M_i(t)} \text{ where } F_i^d = \sum_{j=1, j\neq i}^{N} \text{rand}_i F_{ij}^d(t) \]

\[ F_{ij}^d(t) = G(t) \times (M_{pi}(t) \times M_{aj}(t)) / (R_{ij}(t) + \varepsilon) \times |x_{jd}(t) - x_{id}(t)| \]

6) Updating the particle velocity and its position:
Velocity and position of an object are updated as:

\[ V_i^d(t + 1) = \text{rand}_i \times V_i^d(t) + a_i^d(t) \]

\[ X_i^d(t + 1) = X_i^d(t) + V_i^d(t + 1) \]
B. Parameters of GSA
The performance of a GSA is basically based on two parameters, i.e. swarm size and gravitational constant. Swarm size defines the number of agents that are participating in the function whereas the gravitational constant (G) defines the speed at which solutions change their location in the solution space. In order to achieve better performance, the optimal values for these parameters need to be worked out. Researchers have applied fuzzy logic to control the effectiveness of the parameters of GSA [5] [6] [7]. Initially, a GSA was devised for function optimization but later it has been redesigned for other optimization problems like binary optimization problem [8], multi-objective problems [9], discrete optimization problem [10] and multi-modal problems [11].

C. Performance Issues in GSA
Although GSAs are effective optimization algorithm, yet, like other meta-heuristic techniques, its performance is affected due to its premature convergence, slow convergence rate and long computation time [10]. Researchers have applied several modifications in order to cope with these issues. Operators like Disruption operator [12] and black hole operator [13] have been introduced in GSA to achieve a better balance of exploration and exploitation to avoid premature convergence. Yu Zhang et al. (2012) have proposed immunity based gravitational search algorithm [14]. In this approach, mechanisms of antibody diversity are used to raise the diversity of agents and the characteristic of immunity memory is used to enhance the solution quality and convergence speed of GSA. The concept of Non-dominated Sorting has been introduced in GSA to update the gravitational acceleration of the particles. To promote diversity in GSA, sign and mutations reordering have been used [15]. Exploitation of GSAs has been improved by accelerating the masses towards the best mass (gbest) [16]. A new version of GSA has been implemented which is based on quantum mechanics in which mass of an object is activated by waveform instead of position and velocity[17] [18]. This algorithm suffers from diversity loss problem in collecting the masses of objects. The problem of diversity loss has been removed by an improved version of QGSA[19]. To improve the speed a GSA, a multi-agent based GSA has been proposed by Nanji et al. [20].

D. GSA Hybridization
In order to enhance the capabilities of GSA, it has been hybridized with other conventional and swarm-based algorithms. A hybridized technique named as PSOGSA[21] has been developed in which exploration ability of PSO is integrated with exploration ability of GSA. It improves the convergence speed of GSA and PSO. An adaptive stochastic search (ASS) method has been embedded with GSA in which GSA is used as the global search and ASS is used as the local search. This hybrid algorithm was able to achieve a good convergence characteristic beside its ability to escape local optima [22]. Aizhu Zhang et al. (2015) have proposed a method that hybridized genetic algorithm and Gravitational Search Algorithm. In this crossover and mutation operators of genetic algorithm is used to help the gravitational search algorithm to jump out from local optima [23]. Biogeography Based Optimization (BBO) algorithm has been hybridized with GSA to solve the problem of local population range and to optimize fitness function of GSA [24]. Kepler algorithm has been hybridized with GSA to enhance its performance [25].

IV. GSA IN CLASSIFICATION
Classification is a task of assigning the category or class to unknown data samples by building a model that describes a mapping between the predicting attributes and the class labels of the training data. Many techniques and algorithms have been developed for classification. But each technique has several pros and cons associated with it. This section provides a review of the applications of GSA in the domain of classification.
Hossein Askari and Seyed-Hamid Zahiri (2012) have proposed a GSA based classifier named as intelligent GSA (IGSA) classifier. In this method, a fuzzy system has been used to update the values of the effective parameters, i.e. swarm size and gravitational coefficient. The intelligent GSA is used to find the optimum set of hyper planes from the feature space to distinguish regions so that unknown data set can be classified[6].

A. GSA for discovering classification rules
S. H. ZAHIRI (2012) has proposed a Fuzzy GSA (FGSA) classifier also called fuzzy GSA miner [26] for the discovery of classification rules. In this method, first, the parameter of GSA i.e. gravitational coefficient and the number of effective objects is controlled by a fuzzy controller. Then the improved GSA, called FGSA, is applied for discovery of classification rules [26].

B. GSA for instance-based classification
GSA has been used as a global search method to find the best prototypes [27] and for prototype generation [28] [29] in instance-based classification. Modified version of GSA, Quantum-inspired Binary Gravitational Search Algorithm (QBGSA) have been used with K-nearest neighbor (KNN) classification, where QBGSA is used to find optimum feature subset for the KNN classifier [30] [31]. Jamshidi et al.(2014) have proposed a classifier named as gsaINknn...
SA has been united with Intervals Number K-Nearest Neighbor classifier which is also called the granular extension of conventional KNN.

C. GSA for SVM classification
The classification accuracy of the SVM-based classifier depends on its kernel function and its parameter values. GSA has been widely used to find optimal parameter values in order to improve classification accuracy of SVM [33] [34] [35] [36].

D. Classification technique based on GSA
Researchers have developed classification techniques based on the concept of gravitation and gravitational search algorithm. A data Classification algorithm based on the data gravitation has been developed in which a data sample is classified by comparing the data gravitational force applied by different data classes on that data instance. A data instance will be assigned to the class which applies the largest gravitation force on the instance. This algorithm has a limitation that it suffers from the imbalance of data gravitation [37]. The problem of the imbalance data has been handled by increasing and decreasing the gravitational field of the minority and majority classes. Amplified gravitation coefficient (AGC) that contains class imbalance information has been used for this purpose [38]. Weighted data gravitation classification [39] has also been used to handle the imbalance data problem. In this work, a matrix of weights that describe the importance of each attribute in the classification of each class is used to avoid the imbalance data problem. Shafigh et al. (2013) have suggested a Gravitation based classification algorithm which is based on gravitational potential energy. In this algorithm optimum value of gravitational potential energy of particles is used to find equilibrium condition of classifier. This equilibrium condition of classifier is modeled as a classifier line to differentiate one group of particles from other. However, this algorithm is not suitable for classification of large datasets and multi-class problem [40].

GSA has also been applied in the field of text and image classification. Wan et al. (2016) have applied GSA in texture image classification. In this approach, a GSA has been used to find the optimal “tuned” mask and that tuned mask has been employed for extraction of texture feature [41]. Fatemeh Shirazi and Esmat Rashedi (2016) have defined an approach that is the combination of KNN classifier and GSA. It has been applied for detection of breast cancer tumor [42]. GSAs have been applied to solve several problems related to classification. It has performed well with imbalanced datasets, binary class problems, and multi-class problems. We have observed that most of the GSAs have been applied on continuous datasets. Classification of categorical data using GSA is an under-explored area of research.

V. GSA IN CLUSTERING
Clustering refers to grouping objects based on their characteristics so that the objects within a group have same characteristics and different from the objects in other groups. It is used to classify objects when class label information is not provided. Researchers have applied the various nature-inspired algorithms for clustering. This section highlights the application of Gravitational Search Algorithms in clustering. A number of applications of GSA exist in the related research literature. A data clustering algorithm based on GSA has been proposed in which GSA is used for finding the optimal center for clusters [43]. A GSA has been used with K-means clustering algorithm. In this algorithm, First K-means is used as initial population generator for GSA to improve the convergence speed of GSA then the improved GSA is used with K-means to improve cluster quality [44]. A GSA has been hybridized with heuristic search where GSA was used to produce initial solutions and the heuristic search was used to improve the initial solutions for clustering problem [45]. A GSA has also been applied to K-harmonic mean clustering algorithm for complicated clustering problems. It has shown efficient result but it requires long running time [46]. Further, a GSA has been designed to extract highly correlated bi-cluster of web usage data [47]. A data clustering algorithm based on Grouping Gravitational Search Algorithm has been developed which reduced the redundancy in clustering [43]. A clustering algorithm based on Gravitational Search Algorithm and Genetic Operators has also been suggested in which a GSA has been used to find optimal clusters and genetic operators is used to improve the performance of the GSA [49]. A GSA has been applied to find optimum clusters centers for fuzzy c-means clustering algorithm[50]. Han et al. (2017) have applied an improved version of GSA called Bird Flock Gravitational Search Algorithm (BFGSA) for data clustering problem [51]. Another GSA has been mixed with Big Bang Big Crunch (BBBC) algorithm where GSA is used to find the optimal locations of cluster center and the BBBC algorithm is utilized to diversify the population [52].

VI. GSA IN ASSOCIATION RULE MINING
Association rule mining deals with the discovery of association rules that specifies association and correlation among a set of data items. It deals with the discovery of frequent pattern from a data set to define the interesting associations and correlations within data. Association rules depend on the mined frequent patterns. A GSA based association rule
mining algorithm has been devised that mines strong association rules without generating frequent itemsets [53]. Finding association rules from numeric data sets is complex and hard. A GSA has been designed for mining association rules from the numeric databases [54].

VI. GSA IN FEATURE SELECTION

Feature selection is the process of selecting a subset of features that are relevant for a particular problem. It deals with reducing the size of search space by eliminating irrelevant and redundant feature [2]. Feature Selection helps in improving the predictive accuracy, reducing the complexity and increases the comprehensibility of the knowledge discovered. GSAs have been widely used in feature selection to improve the classification accuracy [30] [31] [33] [34] [35] [36] [42] [55] [56]. J. P. Papa et al. (2011) have proposed a feature selection method based on Gravitational Search Algorithm (GSA) and optimum-Path Forest (OPF). In this method accuracy of OPF is used as the fitness function for the GSA algorithm [57]. Vijay Kumar et al. (2015) have proposed a GSA based automatic feature selection method [58]. This method is efficient for optimal feature selection from high dimensional data.

VII. CONCLUSION

This paper presents a survey of the gravitational search algorithm, its variants, its hybridization with other techniques and its applications in the domain of data mining. From the survey, we can conclude that gravitational search based algorithms have earned a place in the domain of data mining due to their ability to find nearly global optimal solutions in the problems related to large and complex search spaces. This survey shows that GSAs have been widely used in the domain of classification, clustering and feature selection. The research area for association rule mining based on gravitational search algorithm is under-explored. Further, GSAs have shown promising results on numeric data for almost all the data mining tasks. However, applications of GSA for classifying or clustering the categorical data are very rare.

REFERENCES


