# MULTI-OBJECTIVE HYBRID ALGORITHM FOR THE CLASSIFICATION OF IMBALANCED DATASETS

**SANA SAEED**

**UNIVERSITI SAINS MALAYSIA**

**2019**

# MULTI-OBJECTIVE HYBRID ALGORITHM FOR THE CLASSIFICATION OF IMBALANCED DATASETS

**by**

**SANA SAEED**

**Thesis submitted in fulfilment of the requirements**
**for the degree of**
**Doctor of Philosophy**

**March 2019**

# ACKNOWLEDGEMENT

I am thankful to Allah Almighty for all His blessings. This research work could not be possible without the help of Allah Almighty. It is my pleasure to acknowledge the roles of all individuals who were instrumental for completing my Ph.D. research. Many thanks go to my kind and nice co-supervisor Assoc. Prof. Dr. Ong Hong Choon. His guidance and encouragement always boosted me to learn and perform well. I am grateful to him. I am also thankful to my supervisor Assoc. Prof. Dr. Saratha Sathasivam for her guidance. I would like to acknowledge University of the Punjab (PU) for awarding me a scholarship for Ph.D. at Universiti Sains Malaysia (USM). I am also thankful to USM for giving me this opportunity to study in a peaceful and cooperative environment. My deepest appreciation goes to all my family members especially my parents, my husband and my sons for their patience, support, and understanding. At last, I am thankful to all my colleagues and friends for their encouragement.

Sana Saeed

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| *ABC* | Artificial Bee Colony |
| *ACO* | Ant Colony Optimization |
| *AUC* | Area Under Curve |
| $Avg-time$ | Average Running Time |
| *Chisquare* | Chi Square Distribution |
| $CMA-es$ | Covariance Matrix Adaptation Evolution strategy |
| *CS* | Cuckoo Search |
| *CSCMAES* | Hybrid Algorithm Based on CS and CMA-es |
| *CT* | Cart Tree |
| *EA* | Evolutionary Algorithms |
| *ES* | Evolutionary Strategies |
| *FA* | Firefly Algorithm |
| *IR* | Imbalance Ratio |
| *KNN* | K Nearest Neighbors |
| *Mean* | Average Best Value |
| *ML* | Machine Learning |
| *Maxiter* | Maximum Number of Iterations |
| *mkl* | Multiple Kernel Learning |
| *MO* | Multi-Objective |

| | |
|---|---|
| *MOHA* | Multi-Objective Hybrid Algorithm |
| *Mvlognorm* | Multivariate Log Normal Distribution |
| *Mvn* | Multivariate Normal Distribution |
| *Mvt* | Multivariate t Distribution |
| *NB* | Naive Bayes |
| $Niter$ | Current Number of Iterations |
| *PSO* | Particle Swarm Optimization |
| *Sen* | Sensitivity |
| *SDP* | Semidefinite Programming |
| $SMOTE$ | Synthetic Minority Over-Sampling Algorithm |
| *SPEA* | Strength Pareto Evolutionary Algorithm |
| *Std* | Standard Deviation |
| $SVM$ | Support Vector Machines |
| $SVs$ | Support Vectors |
| *SI* | Swarm Intelligence |

# LIST OF SYMBOLS

| | |
|---|---|
| $\alpha$ | Choice of random direction or step size scaling factor |
| $\alpha_{min}$ | Minimum value of alpha |
| $\alpha_{max}$ | Maximum value of alpha |
| $Cov$ | Covariance matrix |
| $Insig$ | Insignificant |
| $m$ | Weighted means generated by CMA-es |
| $n$ | Population size |
| $n_1$ | Class size of the minority class |
| $n_2$ | Class size of the majority class |
| $N_d$ | Number of dimensions of the design variable or number of solutions |
| $P_a$ | Probability of discovering an egg by the host bird |
| $p_{est}$ | Estimated probability |
| $P_{min}$ | Minimum value of the probability of discovering an egg by the host bird |
| $P_{max}$ | Maximum value of the probability of discovering an egg by the host bird |
| $P$ | Total number of Pareto points |
| $p_k$ | Current Pareto point |
| $Sig$ | Significant |
| $x$ | Design variable (solutions) |

| | |
|---|---|
| $X$ | New solution |
| $X_{CS}$ | Best solution by Cuckoo Search |

**ALGORITMA HIBRID MULTI-OBJEKTIF UNTUK KLASIFIKASI SET DATA TAK SEIMBANG**

## ABSTRAK

Klasifikasi set data tidak seimbang kekal menjadi isu penting dalam perlombongan data dan bidang pembelajaran berkomputer. Penyelidikan ini, mencadangkan suatu idea baru berdasarkan pengoptimuman untuk mengendalikan set data tidak seimbang. Suatu algoritma hibrid adaptasi diri baru (CSCMAES) diperkenalkan untuk pengoptimuman. Algoritma hibrid ini berasaskan pada dua algoritma metaheuristik yang terkenal: Carian cuckoo (CS) dan strategi evolusi adaptasi matriks kovarians (CMA-es). Untuk penumpuan pantas dan untuk prosedur carian yang efisien, adaptasi diri dalam parameter algoritma hibrid dicadangkan. Keberkesanan algoritma ini diuji dengan masalah fungsi ujian tanpa kekangan dan dengan kekangan melalui kajian simulasi. Daripada kajian simulasi, adalah ditunjukkan bahawa CSCMAES dilakukan dengan baik pada setiap fungsi ujian dan menghasilkan nilai terbaik dengan sisihan piawai minimum dan dengan penumpuan lebih cepat. Selepas itu, suatu algoritma hibrid multi-objektif (MOHA), iaitu lanjutan daripada algoritma hibrid adaptasi kendiri dicadangkan dan diuji pada fungsi ujian multi-objektif (MO) yang telah ditetapkan. MOHA yang dicadangkan mendapat keputusan, baik dalam fungsi ujian ini. Suatu metodologi baru dibentangkan untuk klasifikasi set data tidak seimbang. Idea utama metodologi ini adalah untuk menganggarkan kebarangkalian untuk setiap kes dalam kedua-dua kelas secara berasingan. Untuk tujuan ini, taburan normal digunakan pada setiap kelas. Parameter taburan ini dioptimumkan dengan aplikasi MOHA yang dicadankan. Prestasi cekap metodologi yang dicadangkan ini diperhatikan dengan bantuan kajian eksperimental pada tiga jenis set data; data simulasi, data bersempadan bising dan set data

tidak seimbang sebenar digunakan. Tambahan pula, prestasi mesin vektor sokongan (SVM) yang dipertingkatkan dikaji dengan menggunakan algoritma prapemprosesan dan pengoptimuman. Di samping itu, prestasi SVM juga dikaji dengan pembelajaran kernel berbilang (mkl) menggunakan algoritma prapemprosesan dan pengoptimuman.

# MULTI-OBJECTIVE HYBRID ALGORITHM FOR THE CLASSIFICATION OF IMBALANCED DATASETS

## ABSTRACT

Classification of imbalanced datasets remained a significant issue in data mining and machine learning (ML) fields. This research work proposed a new idea based on the optimization for handling the imbalanced datasets. A new self-adaptive hybrid algorithm (CSCMAES) is introduced for optimization. The proposed algorithm is grounded on the two famous metaheuristic algorithms: cuckoo search (CS) and covariance matrix adaptation evolution strategy (CMA-es). For its fast convergence and for its efficient search procedure, the self-adaptation is proposed in the parameters of the proposed hybrid algorithm. The effectiveness of this algorithm is verified by applying it on the unconstrained and constrained test functions through a simulation study. From the simulation study, it is shown that CSCMAES performed very well on each test function and produced the best values with minimum standard deviation and with faster convergence. Thereafter, a multi-objective hybrid algorithm (MOHA), an extension of the self-adaptive hybrid algorithm is proposed and tested on the established multi-objective (MO) test functions. The proposed MOHA performed very well on these test functions. A new methodology is presented for the classification of the imbalanced datasets. The key idea of this methodology is to estimate the probabilities for each case in both classes separately. For this purpose, the normal distributions are applied to each class. The parameters of this distribution are optimized by applying the proposed MOHA. An efficient performance of this proposed methodology is observed with the help of an experimental study in which three types of datasets; simulated datasets, noisy borderline datasets and real-life imbalanced datasets are engaged. Fur-

thermore, an improved performance of support vector machines (SVM) is studied by using the preprocessing algorithm and optimization. In addition, the performance of SVM is also studied with multiple kernel learning (mkl) by applying the preprocessing algorithm and optimization.

# CHAPTER 1

# INTRODUCTION

## 1.1 General Introduction

This chapter will provide a general idea of this research work, in which, research area and the elementary methodologies are introduced. Research problems that are investigated in this study and related questions will be discussed. The objectives of this research work are also defined in this chapter. A complete layout of this thesis is presented in the last paragraph of this chapter.

### 1.1.1 Optimization

Optimization relates to various fields with a wide variety of applications. Human beings always have some constraints on their resources such as time constraint and financial constraints. Therefore, its significance cannot be denied. Optimization problems can be partitioned into two categories based on the type of the decision variables i.e. discrete optimization or continuous optimization. The solution to a problem is usually found by an efficient optimization procedure either by minimization of the cost function or by the maximization of the performance measures. Efficient optimization mostly uses the derivative information obtained from the cost function based on the design variables. However, it is difficult to obtain the accurate information in many real-life situations or the evaluation of cost function with the existing methodologies is too expensive.

Last decade has seen a rapid growth of optimization procedure which can work without the derivative information of the cost function with the fastest convergence. These procedures are usually termed as derivative-free optimization methods (Kramer et al., 2011 ; Rios & Sahinidis, 2013). Tools for finding the rapid and accurate solution of optimization problems are algorithms. Most conventional and classical algorithms produced suboptimal results particularly for multimodal and high dimensional problems due to their deterministic nature (Rakhshani & Rahati, 2017; Yang, 2010). After the development and modifications of classical algorithms, enhanced nature-inspired algorithms are introduced by the researchers which are capable enough to overcome all the inadequacies of the classical algorithms. Nature-inspired algorithms have been developed by getting inspiration from nature. These algorithms are stochastic and derivative-free in nature (Fister Jr et al., 2013; Mlakar & Fister, 2016).

### 1.1.2 Nature-Inspired Metaheuristic Algorithms

Nature-inspired metaheuristics algorithms, a type of stochastic algorithms have the capability of providing the good quality solutions of problems using both randomness and local search (Talbi, 2002). Lots of work have been done on nature-inspired metaheuristics by the researchers. Swarm intelligence (SI) and Evolutionary algorithms (EA) are the two major branches of these algorithms (Gendreau & Potvin, 2010; Lones, 2014). The key idea behind EA is that only those entities of a population which meet certain selection criteria are kept and the rest are discarded (Fister Jr et al., 2013). In this way the population will converge to those entities that fulfill the selection criteria (Parrill, 2000). SI is the combined intelligence of clusters of all agents. Algorithms based on SI must be flexible to internal and external changes, should be

robust, distributed and self-organized (Chu et al., 2011; Deepa & Senthilkumar, 2016).

### 1.1.3 Intensification and Diversification

Metaheuristic algorithm is considered to be an efficient device to produce the optimal solutions within an adequate time for the complex optimization tasks. To find the good and reasonable solutions for the complex problems, a metaheuristic algorithm must hold two characteristics: (1) generate the efficient solutions which should be more effective than the existing solutions by searching the whole area where the global solutions can be found (2) to be able to escape from the local optimum. The combination of these two characteristics (intensification and diversification or exploration and exploitation) are highly demanded (Blum & Roli, 2003; Lozano & García-Martínez, 2010).

Exploration is usually performed by the randomization procedure which enables an algorithm to search globally and don't get stuck in local optimum. The search procedures based on randomization can also be engaged in a local search for the whole space near the best solution if the steps are restricted to the local area. The randomization can search the space globally for the large steps. Whereas, exploitation uses the information of the local space and produces the local optimum. Exploitation increases the convergence speed of metaheuristics. However, exploration decreases the convergence rate of an algorithm and reduces its efficiency. A fine-tuning of these two tools of metaheuristic algorithms can enhance its efficiency (Yang et al., 2014).

### 1.1.4 Hybrid Algorithms

The idea of combining two or more algorithms is also now gaining popularity day by day due to their optimal results. Taking the benefits from two or more algorithms, the newly proposed algorithms are usually recognized as hybrid algorithms (Cung et al., 2006). Many studies have been conducted on hybrid algorithms. Few of them used metaheuristics with local search for hybridization and few utilized only nature-inspired algorithms i.e. evolutionary and swarm algorithms (Bianchi et al., 2006; Blum et al., 2011). Most of the hybrid algorithms are problem specific and are being proposed to solve specific problems by taking the advantage of optimization abilities of two or more algorithms (Abdullah et al., 2012; Chiarandini et al., 2006; Vidal et al., 2015).

For ML and data mining fields, particularly for the classification of imbalanced datasets, different hybrid algorithms have been introduced by different authors. However, most of the algorithms are only based on the combinations of ML approaches and metaheuristic algorithms. Few of them are discussed in Chapter 2. For the classification of imbalanced datasets, the limited material is available on the nature-inspired hybrid algorithms.

Two metaheuristic algorithms, CS and CMA-es are combined in this study, to build a newly proposed hybrid algorithm. CS is grounded on the characteristics of swarms whereas CMA-es are based on the evolution strategies. During this research work, the joint efficiency of these algorithms will be utilized to produce a new hybrid algorithm. Afterwards, this newly proposed algorithm will be extended to MOHA to solve MO problems.

### 1.1.5 Classification of Imbalanced Datasets

The handling of imbalanced datasets is a significant problem in ML and data mining. These types of datasets can be originated in many real-life applications, for example, recognition of fake telephone calls, text classification, in marketing and in the medical field, etc (Nikulin & McLachlan, 2009; Phua et al., 2004; Zheng et al., 2004). These types of datasets have an imbalance among classes i.e. one class has much more instances than the other classes. For binary datasets, a class having many instances is recognized as a majority (negative) class and the other one is termed as a minority (positive) class. Traditional classifiers are not capable enough to handle this issue as they show their biased behavior for the majority classes (Chawla et al., 2004; He et al., 2008; He & Garcia, 2009).

The problems of imbalanced datasets also increased if the datasets contain noise. Noise can originate in imbalanced datasets due to many reasons, for example, from incorrect labeling, because of an inadequate number of examples in the data collection phase, and from data preparation stages. Many ML algorithms are badly affected by the noise. However, this problem gets worse if the dataset contains the imbalance problem also. The reason is that many standard ML algorithms usually consider minority class as the noise of the dataset (Garcia et al., 2012; Weiss, 2004).

Borderline instances or examples are another issues of noisy imbalanced datasets. These examples are usually positioned in the neighboring area of the class boundaries where most of them are overlying. This is again a challenging task for the researchers in data mining and ML.

The significant reasons behind the reduced performance of ML algorithms include the ignorance of class-wise efficiency and considering only the overall performance, the assumption of the equally distributed data among all classes and considering the equal cost of the loss among classes (Kumar & Sheshadri, 2012).

Among all methods, the distribution of equal error cost to all classes is the foremost disadvantage of the existing classifiers because of the reason that for many practical applications, misclassification of examples in a dataset may create a different problem for different classes. For example, the wrong diagnosis in medical field, i.e misclassifying the cancerous cells may cause an adverse health risk. For this purpose, to tackle this serious issue, many methodologies have been suggested by the authors.

Different types of sampling methods are presented for handling the imbalanced datasets. These methods introduced the artificially generated examples with the help of resampling. These methods are usually recognized as the pre-processing methods (Chawla et al., 2002; He et al., 2008; Van Hulse & Khoshgoftaar, 2009).

Another way to intelligently tackle the imbalanced datasets is the feature selection which is also very popular among the researchers. The feature selection methodologies have been applied in combination with different classifiers or with different sampling methodologies for these types of datasets which can be studied from the existing literature. For example, the concurrently engaging the backward elimination feature selection method and SVM for imbalanced high dimensional datasets and the use of undersampling feature selection and SVM on the skewed datasets (Al-Shahib et al., 2005; Maldonado et al., 2014).

Because of the successful applications of metaheuristic algorithms and computationally intelligent techniques for different real-world optimization problems particularly, in ML and data mining, no one can ignore their important role (Duval & Hao, 2009; Yang et al., 2014). Many studies have been conducted using metaheuristics algorithms and proved a significant performance for handling classification issues. For example, to deal with imbalanced datasets, the proposal of creating the artificial examples for the minority class using a genetic algorithm (GA) had been introduced by Beckmann et al. (2011). A hybrid system, showing the use of SI particularly particle swarm optimization (PSO) with multiple classifiers and evaluation metrics had been introduced by Yang (2009).

From the above discussion on the imbalanced datasets and different proposed methods for tackling them, it can be said that this is a significant issue in data mining and ML. Keeping in view the significance of this issue, this study is going to introduce an efficient methodology by using the MOHA.

## 1.2 Research Questions

The basic research question to be addressed in this study is how to classify imbalanced datasets using the characteristics of metaheuristics algorithms. A bridge between these two fields, data mining, and optimization algorithms, should be built so that the data mining issues should be studied by taking the advantage of the latest metaheuristics optimization algorithms. For this purpose, the first research question is how to form a new hybrid algorithm which would hold the features of two nature-inspired algorithms. One of them is the CS which grips the characteristics of SI and another one

is the CMA-es that holds the properties of EA. The next question is whether this newly formed self-adaptive hybrid algorithm is capable enough to compete with the already existing algorithms. The third important question attached to this newly formed algorithm is the convergence behavior of this algorithm. How can we build its MO version? To study the validation of MO algorithm is another research question for this study.

The one and most significant research question is how to propose a new methodology for the given classification task of imbalanced datasets using the MOHA. The behaviors of different datasets i.e simulated, synthetic and real life datasets with the proposed methodology will be investigated in this study. Besides these main research questions, the other research questions are what are the applications of the proposed optimization algorithm in the ML field. For example, one of the applications of optimization algorithm can be seen for the parameters selection of existing classifiers via random search and in the field of kernel learning. Therefore, another research problem for this research work is, how to improve the performance of SVM, for the occurrence of imbalances in the datasets. To improve the performance of SVM for imbalanced, noisy and borderline datasets, an intelligent methodology would be introduced. How to apply the proposed optimization algorithm in the field of kernel learning with the weights optimization problem is another research question.

## 1.3 Research Objectives

1. To propose a new hybrid algorithm by using the combined efficiencies of CS and CMA-es. Self-adaptation in parameters will be introduced for this purpose. Implementation of this proposed algorithm will be done on unconstrained and constrained test functions.

2. To propose a MOHA, the self-adaptive hybrid algorithm will be extended. In the classification task of noisy borderline and imbalanced datasets, a new methodology will be proposed. For this purpose, the two significant research areas, optimization, and ML will be combined by using MOHA and the generative classifier (normal distribution).

3. For the extensions and applications of this work, the self-adaptive hybrid algorithm will be applied in ML field. For this given task, the proposed self-adaptive hybrid algorithm will be used with the supervised classifier.

4. To study the improved performance of SVM for the given classification task of imbalanced datasets by using a preprocessing oversampling algorithm and optimization, a new methodology will be proposed. A comprehensive experimental study will be conducted by using the synthetic noisy, borderline and real imbalanced datasets, to confirm the validity of the proposed methodology.

5. To observe the performance of SVM with mkl for synthetic noisy, borderline and real imbalanced datasets. Another methodology will be introduced by using an oversampling algorithm and optimization for the kernel weights and the parameters. An experimental study will be conducted with more than one kernel functions.

## 1.4 Arrangement of Thesis

This thesis has seven chapters including the first chapter as an introduction. In Chapter 2, all the relevant studies based on optimization and the classification of imbalanced datasets will be discussed. The performance of SVM discussed by different authors by applying different methods will also be presented. The newly proposed self-adaptive hybrid algorithm will be presented in Chapter 3 and thereafter the simulation study on this new algorithm and comparisons with other algorithms will be presented and discussed in Chapter 4. Chapter 5 is based on MOHA followed by a newly proposed methodology for the classification of imbalanced datasets. Chapter 6 will show the extensions and applications of this work. Finally, a summary, some extensions for this study and the future directions will be discussed in Chapter 7.

# CHAPTER 2

# REVIEW OF LITERATURE

In this chapter, the literature related to the objectives of this study will be reviewed. All the related studies will be discussed in different sections with respect to different aspects of the study, starting from nature-inspired to mkl. However, to justify the significance of this present research work, the focus of discussion will be on those studies which were conducted for the classification of imbalanced datasets using different optimization techniques.

## 2.1 Studies on Nature-Inspired Metaheuristic Algorithms

Optimization is a way of finding the optimal solution of the problem under consideration from the given potential sets of substitutes following the defined criteria. This method involves the maximization or minimization of a real-valued function by progressively picking the values from a stable predefined range and generating the best values of the given task. The obtained "best" solution means that no other solution is equal to or better than it. The algorithms engaged for the optimization problems can be of deterministic and stochastic. Since the former method comprises substantial and tedious calculations, thus, the later optimization methods are preferred by the investigators (Binitha et al., 2012; Haupt & Haupt, 2004).

In recent years, metaheuristic algorithms inspired by nature are commonly used for answering optimization problems. The acceptance of metaheuristic algorithms in

the field of optimization is because of their fast performance. Although their produced solutions are not optimal but these solutions are valid because they do not take long running time. Intensification and diversification are two essential features of metaheuristics. Finding the best solution by moving around the existing best solution is done by intensification. However, diversification examines the whole area for providing the global optimal solution.

According to Blum and Roli (2003) and Talbi (2002), a good metaheuristic algorithm should hold the balance of the two. Therefore, an algorithm should be quick enough in determining the region of good quality solutions in the whole search area and secondly, it should not spend much time for those regions which do not have a good quality solution or which are already explored.

Many nature inspired metaheuristic algorithms are very popular and are applied by different researchers in various fields requiring high-quality solution in their real-life problems. These algorithms are divided into two major categories (1) SI and (2) EA. The most famous algorithms having the behavior of swarms are PSO, ant colony optimization (ACO), CS, and firefly algorithm (FA) (Blum, 2005; Kennedy, 2011; Yang, 2009; Yang & Deb, 2009).

Evolution based common algorithms include genetic programming (GP), genetic algorithm (GA), evolutionary programming (EP), learning classifiers systems (LCS) and evolution strategies (ES) (Parrill, 2000). Among the class of EA's, evolution strategies based on the Gaussian mutation are popular for parameter optimization. CMA-es is the utmost popular and effective ES among real-parameter optimization for non-

linear problems (Auger et al., 2004; Hansen & Kern, 2004).

## 2.2 Studies on Cuckoo Search and Covariance Matrix Adaptation evolution strategies

This study proposes a new self-adaptive hybrid algorithm based on CS and CMA-es. Therefore, different existing studies related to these two algorithms are presented here which will help us in understanding their significant roles in handling optimization problems in different fields. However, the main focus will be on the implementations of these algorithms for the classification task of imbalanced datasets.

CS is a famous nature-inspired algorithm developed by Yang and Deb (2009). This algorithm is inspired by the attitude of cuckoos. These birds attract others not only by their sounds but also by their hostile behavior of reproduction strategy. The later version of this algorithm was improved by levy flights instead of the simple random walk. Many advancements and modifications of this algorithms have been proposed by different researchers. Modifications for unconstraint optimization, modification using Mantegna levy flights, modification using exchange of information, problem-specific modifications, enhancements using self-adaptation of parameters in the algorithm have been introduced (Li & Yin, 2015; Naik et al., 2015; Nguyen & Vo, 2015; Tuba et al., 2011; Walton et al., 2011).

Fateen and Bonilla-Petriciolet (2014) proposed a simple modification of CS for global optimization using the information obtained by the derivative of the objective function. This modification proved to be consistent and effective for most of the test functions. However, improved performance could not be achieved for the benchmark

test problems.

The use of CS algorithm can be found in different fields and for different types of problems, for example, for multimodal function, engineering optimization, for business problems, design of steel structures and constrained problems (Bulatovic et al., 2014; Cuevas & Reyna-Orta, 2014; Yang & Deb, 2010; Yang et al., 2012). The comprehensive literature reviews by two different authors on CS were done by Fister Jr et al. (2014) and Mohamad et al. (2014).

In the field of ML, different applications of CS can be found. For example, this algorithm is used for feature selection in the classification task by forming its binary versions (Pereira et al., 2014; Rodrigues et al., 2013). Wang et al. (2016) introduced a nearest neighbor CS algorithm (NNCS) with the probabilistic transformation. The author proposed an idea of utilization of nearest neighbors strategy to search new solutions instead of best solution so far. A solution based and fitness based similar metrics were employed for the implementation of nearest neighbor strategy in CS.

Husaini et al. (2016) proposed a modification in CS using Markov Chain Monte Carlo (MCMC) and this modification showed a satisfied performance with respect to the convergence. A recent study proposed by Ismail et al. (2017) presented another modified CS algorithm for solving quadratic assignment problem (QAP). To handle the discrete variable of QAP, the smallest position value rule was introduced. The application of CS for the medical data, remote sensing data, and the cancer data classification could also be found in the different studies proposed by Bhandari et al. (2014), Gunavathi and Premalatha (2015), and Mohapatra et al. (2015).

A significant role of MO algorithms cannot be denied. Therefore, we can also see different MO versions of CS proposed by different authors using varying strategies. The weighted sum approach and the non-dominated sorting were mostly engaged for MOCS (Balasubbareddy et al., 2015; Rani et al., 2014; Yang & Deb, 2013).

As mentioned earlier, CS had been used in ML, particularly for classification tasks. However, the available material is insufficient. Another important issue is that most of the studies used this algorithm are only for feature selection purposes using its different versions. However, a study by Abdualrhman and Padma (2017), proposed a robust and scalable classifier based on CS. This was the first study which used CS algorithm in making a binary classifier directly. According to this author, CS was employed for the class search due to the rapid search characteristics of this algorithm.

Now we will discuss the existing studies on another optimization algorithm used in the formation of our new hybrid algorithm. Before performing hybridization, it is better to view and understand its role in different fields. This algorithm is another nature-inspired algorithm and follows an evolution theory proposed by Darwin. There are different algorithms based on this evolution theory but only CMA-es is used in this study.

CMA-es was introduced by Hansen and Ostermeier (1997). According to Muller et al. (2009), CMA-es is one of the ES and a stochastic iterative technique for continuous parameter optimization. Few developments and modifications were proposed by the authors in different years (Hansen et al., 2003; Hansen & Ostermeier, 2001). Hansen and Kern (2004) evaluated CMA-es by using the multimodal test functions.

The convergence of CMA-es was also studied by Diouane et al. (2015).

A restart strategy of CMA-es with the increased population size at each restart for the global optimization problems was suggested by Auger and Hansen (2005) and Suganthan et al. (2005). The authors evaluated this method on the 25 real-parameter optimization test functions of CEC 2005.

In another study, the effect of small primary population size on IPOP Active CMA-es with mirror mutations was investigated by Brockhoff et al. (2012). Hansen (2008) proposed another idea of combining two-point step size adaptation with CMA-es for large populations. In addition to this combination, a refined formula for the learning rate of the covariance matrix and recombination weights was also suggested.

This algorithm also has some drawbacks in terms of its efficiency. According to Chen et al. (2009), although CMA-es is the famous EA for solving continuous optimization problems, its efficiency is not much admirable but these algorithms have the ability of escaping from the local minima. Therefore, in many studies, different adaptations, and modifications of this algorithms were proposed to cover its shortcomings.

Beyer and Sendhoff (2008) proposed a new adaptation strategy in CMA-es to overcome the drawbacks of these algorithms. The weighted recombination is a way for improving the local search of evolution strategies by making use of available effective information. A ranked based weighted recombination was introduced. The optimal weights were computed for the sphere model and comparisons were made with the strategies without weighted recombination. An extension of this study was made in which the weighted recombination strategy was studied for the parabolic ridge (Arnold,

2005, 2006).

As CMA-es is a continuous optimization algorithm. This algorithm has been applied to the optimization problems of real life. Suominen et al. (2012) used CMA-es for the parameter estimation of complex chemical kinetics. An efficient improved CMA-es for network security situation prediction was proposed by Hu and Qiao (2015). CMA-es was engaged for the total cost minimization of energy and the spinning backup arrangement for a wind thermal power in the study proposed by Reddy et al. (2013). In another study, a multimodal, MO and nonlinear optimal transformer design was proposed. For the optimization of that design, CMA-es was used by minimizing the four objective functions namely purchase cost, total lifetime cost, total mass and total loss individually (Tamilselvi & Baskar, 2014).

The use of CMA-es for MO optimization problems can also be studied from the available literature (Igel et al., 2007). Thereafter, an improved step size adaptation was proposed in the MO version of CMA-es (Vob et al, 2010). However, a limited material is available on CMA-es as a MO optimization algorithm.

After viewing the literature, it can be said that these two algorithms are popular optimization algorithms for continuous test problems. CS, as it is discussed above, is capable enough of generating the efficient results in a minimum time, whereas CMA-es has a few drawbacks in terms of efficiency. Therefore, this study took the inspiration from these two algorithms in terms of combining them and form a new hybrid algorithm which would hold the characteristics of these two optimization algorithms. CS holds the features of swarm intelligence whereas CMA-es holds the properties and con-

cepts of the evolution process. So by hybridizing them, we would be able to combine both the characteristics in one algorithm, which will be capable enough to overcome the shortcomings of CMA-es.

### 2.3 Studies on Hybrid Algorithm of CS and CMA-es

Over the last few years, the interest in hybridization has increased. The hybridization can be done in several ways. The first way is the addition of different components of one algorithm to another algorithm. The control of these schemes is grounded in the impression of recombining the results to acquire new ones. The second way is to form the concern systems considered as the cooperative search. This method consists of exchanging the information through different algorithms in some way. The third possible way is the mixing of approximate (or complete) schemes (Jourdan et al., 2009; Talbi, 2009). In this section, a look at the existing hybrid algorithms based on CS and CMA-es will be taken.

A hybrid algorithm of CMA-es and hybrid differential evolution (HDE) was proposed by Kampf and Robinson (2009). The main focus of the authors was on varying placement of buildings to optimize solar irradiation availability. Kanagaraj et al. (2013) introduced another hybrid algorithm for reliability and redundancy allocation problems. This hybrid algorithm was formed by using CS and GA (CS-GA). By inserting GA operators in standard CS algorithm, an improved exploration and exploitation of the algorithm was achieved.

Feng et al. (2014) introduced another hybridization of CS with Shuffled Frog Leaping algorithm. This hybridization was also done for the real world problems. The re-

searchers applied this proposed algorithm for solving the Knapsack problems. For the efficient constraints handling, in optimization issues by applying the CS algorithm, a hybrid version was introduced by Long et al. (2014). This hybridization was done with a local search technique Solis and Wets method by engaging the lagrangian method for constraints.

A hybrid Kalman CS tracker was suggested by Ljouad et al. (2014). A modified version of CS was combined with Kalman filters. This hybrid enhanced the quality of the initial population and produced better results in terms of computational time.

Prachi and Kaur (2015) introduced another hybridization of two SI algorithms, CS and artificial bee colony (ABC). The proposed algorithm was introduced for an improved and efficient classification of the satellite image.

A hybridized CS algorithm was observed for the cluster analysis, a renowned method in data mining. The algorithm was proposed with the combination of differential evolution algorithm for data clustering. By taking advantage of differential evolution algorithm in CS algorithm, better results in terms of convergence analysis was observed (Bouyer et al., 2015).

Mlakar and Fister (2016), proposed a hybrid self-adaptive CS algorithm. The suggested algorithm was mainly an expansion in the actual scheme of CS with the balancing of the exploration schemes, self-adaptation for the parameters and the linear population reduction.

Two-hybrid algorithms of CS with Nelder Mead method were proposed. Both

algorithms were problem specific, one was for integer programming and another one was for the optimization of the multi solar system (Ali & Tawhid, 2016; Jovanovic et al., 2014).

A comprehensive study of the various versions of the modified CS (MCS) with the strength pareto evolutionary algorithm (SPEA) was conducted for the rectangular arrays. Modification in CS was proposed with the roulette wheel selection operators to select the primary host nests. In the 3D search area, the adaptive inertia mass to regulate the locations, search of the possible finest host nest and the dynamic detection amount to regulate the fraction of the likelihood of discovering the finest host nests were also selected with the help of the proposed modifications. This study also proposed two hybrid algorithms of this modified CS with PSO and hill climbing with SPEA (Rani et al., 2017).

Sun and Gu (2017) proposed a hybrid algorithm of CS to solve the flow shop scheduling task. The hybrid estimation of distribution algorithm was combined with CS to form a new hybrid algorithm. A discrete solution representation method was applied to increase the operation efficiency.

After discussing the literature on hybrid algorithms, it can be perceived that all projected hybrid algorithms were problem specific and introduced for handling these problems in the framework of optimization. Different studies have proposed different hybrid algorithms for the real-life problems. However, the only single study is available on the hybridization of CS and CMA-es algorithm (Rakhshani & Rahati, 2017), in which hybridization was done for intelligent multiple search strategy algo-

rithm (IMSS) with Q learning but a decent point is the further exploration of different ways of forming a new hybrid algorithm. Therefore, this area necessitates a special attention, because of the effective results reported by these nature-inspired metaheuristic algorithms with respect to time, fast convergence, and robustness. The inadequate material and few implementations of hybrid algorithms in ML area and mainly for the classification field of imbalanced datasets again give us a motivation for exploring and proposing a new hybrid algorithm.

## 2.4 Studies on the Classification of Imbalanced Datasets

In data mining, the learning of datasets and predominantly for imbalanced datasets is a substantial matter. Many researchers have studied and addressed this prevalent issue. Various ideas were introduced to overcome this issue. The suggested approaches include sampling approaches, feature selection approach, and the algorithm approaches. The hybrid methods based on these approaches can also be found in the literature. We will discuss here all the possible related studies on the imbalanced datasets, which forced us to think over this problem. However, our focus will be on those studies which used metaheuristics optimization algorithms for the imbalanced datasets.

A thorough material for studying the patterns of imbalanced datasets can be seen in Guo et al. (2008), Maheta and Dabhi (2015), Phung et al. (2009), Witten et al. (2016) and Yang et al. (2009). Different approaches including data imbalance, sampling techniques for handling it, including basic sampling and advanced sampling and algorithm level methodologies were discussed by the authors in the previously mentioned studies.

Batista et al. (2004) conducted a study on different sampling techniques to observe

their performances for the balancing of training datasets. It is shown that oversampling produced better results than the other methods including the undersampling using the area under the curve (AUC). In another study, synthetic oversampling methods for increasing the classification accuracy were applied. Moreover, two alterations to the prior methods namely SLOUPS and OUPS were introduced (Rivera & Xanthopoulos, 2016).

Bach et al. (2017) conducted an experimental study based on the undersampling and oversampling methods for the analysis of highly imbalanced datasets regarding osteoporosis. The objective of the study was to identify a better sampling approach for the modest and ensemble-based classifiers.

The use of Bayesian methods for handling imbalanced datasets was introduced by Maragoudakis et al. (2000). Galar et al. (2012) presented a brief review of different methods including bagging, boosting and hybrid-based methods for imbalanced datasets.

Existing classifiers were also explored by researchers in the presence of imbalanced datasets. For the classification task of imbalanced datasets, Sun et al. (2007) introduced a cost-sensitive boosting. The authors, for this purpose, suggested a cost into the framework of AdaBoost. The cost-sensitive boosting algorithm was also explored for the weighting schemes related to different types of samples.

A comprehensive study of the performance of k nearest neighbors (KNN) for imbalanced and overlapping datasets can be studied from the literature (Garcia et al., 2008). Applications of the cost-sensitive techniques for imbalanced datasets are also

common. For example, Thai-Nghe et al. (2010) used cost-sensitive learning methods and resampling methods for handling these types of datasets. From their two proposed methods, the first one was a combination of resampling method and SVM and the second was grounded on the optimization of the cost matrix.

Boonchuay et al. (2017) suggested the use of entropy for the minority class of imbalanced datasets. The improved classification results were observed for these datasets by using the decision tree algorithm with the proposed minority class entropy.

In another study, for the imbalanced training datasets, a cost-sensitive margin distribution learning was developed and introduced a large cost sensitive margin distribution machine (LCSDM). The proposed method gradually increased the marginal distribution of the positive class to obtain the balanced classification results (Cheng et al., 2017).

Model-based approaches are also common for imbalanced data learning. The use of combinative classifiers can be justified from the literature. An effective handling method for the imbalanced datasets was introduced with a model fusion approach by incorporating the discriminative classifiers (cSVM) and the generative classifiers (GMM) (He et al., 2015).

Chen et al. (2004) applied the random forest for handling the imbalanced datasets in two different ways called weighted random forest and balanced random forest. The weighted random forest assigned comparatively more weights on the positive class. Whereas, balanced random forest entailed the idea of joining downsampling (from the majority class) and ensemble learning. With the large imbalanced datasets, the second

realization of random forest, i.e. the balanced random forest was observed to be more effective.

Because of the rising popularity of computational intelligence and metaheuristic algorithms in various fields, for example, in computer science, artificial intelligence, ML and data mining, frequent use of these procedures and methods can be seen for the classification of imbalanced datasets. For example, the learning classifiers systems also called evolutionary online rule-based systems were examined to prove their abilities of mining sufficient amount of information from the imbalanced datasets. It is shown that learning classifier systems were capable enough to extract the sufficient information from the imbalanced datasets (Orriols-Puig & Bernadó-Mansilla, 2009).

Milare et al. (2010) introduced a hybrid approach using evolutionary algorithms to learn the imbalanced classes. The proposed rule sets are combined with an evolutionary algorithm to build a new classifier. Ducange et al. (2010) introduced an idea of applying MO genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. The potentially optimal classifiers in the projection of Pareto front approximation were selected by using receiving operating curve convex hull method. A reduced classification cost was obtained by using the proposed scheme.

For the imbalanced and borderline datasets, the hierarchical genetic fuzzy system based on genetic programming was introduced. The results were verified with the help of a statistical analysis of an experimental study (Lopez et al., 2013). To enhance the classification ability of the classifiers for cancer diagnosis, the feature extraction method based on genetic programming was suggested by Moreno-Torres et al. (2013).