The Impact of Data Mining and SaaS-Cloud Computing: A Review

Abir Achache [1,*], Abdelhalim Baaziz [1], Toufik Sari [1]

[1] *LabGED Laboratory, Computer Science Department, Badji Mokhtar University, Annaba, Algeria*

***Abstract*— Cloud Computing has emerged as a powerful paradigm that has successfully dominated network services in many application areas and significantly transformed the IT industry. In the Cloud Computing, all resources are available as services and accessible via the Internet. Software as a Service is considered as the king of service delivery models that enable end-users to access to any software or application as a service via the Internet, without local installation. Over the past decade, this model has been widely adopted by many organizations and individuals, leading to the production and accumulation of a huge amount of data stored in the Cloud from distributed nodes that must be recovered very efficiently. Software as a Service providers must be able to manage this data successfully, evaluate and improve the quality of their solutions in order to provide reliable and efficient services to users of Cloud services. On the other hand, Data Mining is a current trend in the field of data treatment that allows the extraction of useful and meaningful information from raw data. The integration of Data Mining techniques into Cloud Computing – Software as a Service - has become commonplace and can support the adequate provision of services by providing agility and rapid access to technology. This article introduces the basic concepts of Data Mining and Cloud Computing first, and discusses the coupling of the two paradigms. Then, it describes how Data Mining can be used and integrated to improve Software as a Service services in the Cloud. Finally, it reviews relevant and important research in this area.**

***Keywords*— Classification, Clustering, Cloud Computing, Data Mining, KDD, Software as a Service.**

I. INTRODUCTION

Over the last few years, the Internet has succeeded in reserving considerable importance as an indispensable tool in our personal and professional lives, which is reflected in the enormous increase in its users. According to "Digital Around The World" in January 2019, 57% of the world's population now uses the Internet (4.388 billion people) as an innovative computer communication system that has become an important place every day. This evolution is associated with the emergence of a variety of technologies among Cloud Computing. This latter is considered as one of the most revolutionary concepts of 21st century information technology that has realized the dream of utility computing and is increasingly penetrating into all business and scientific areas.

Cloud Computing is the new trend in hosting and delivering Internet services that rely on server Clouds to manage tasks. It is a change in the service-oriented IT environment where IT infrastructure and solutions are provided as a service to users on demand via the Internet [1, 2]. The National Institute of Standards and Technology has defined Cloud Computing as a computing paradigm that involves the use of hardware and software resources (e.g. networks, servers, storage, applications and services) provided as a service in a dynamic and personalized way to several end-users over the Internet [3]. IT resources can be provisioned, liberated dynamically by users as an on-demand service and consumed according to their requirements. Cloud Computing is a distributed computing platform in which resources are shared among several users and are accessible from anywhere and at any time as long as an Internet connection exists. Cloud computing adopts a pay-per-use pricing system, which means that users pay only for what they actually use and for the time they need [1, 5, 6, 7, 8, 9, 10]. As required, Cloud Computing can be deployed under a private, public, community or hybrid model [1, 4, 5, 9]. It adopts a service-oriented approach where resources are offered as a service on demand according to three models: IaaS (infrastructure as a service), PaaS (platform as a service) and SaaS (software as a service) [1, 6, 11, 12, 13]. The latter is considered the leader of Cloud service delivery models. Software as a service (SaaS) sometimes referred to as software on demand is a software deployment model where an application is hosted as a service provided to customers over the Internet via a pay-per-use or free payment, which allows users to get rid of the burdens of the local installations and executions of applications [1, 6, 7, 10, 11, 14]. The favorable context of SaaS has permitted its adoption by organizations and individuals in a drastic way. In a similar instance, Cloud Computing under its various forms is becoming one of the trendy words of the next industry; its use has gained popularity due to: its mobility, high availability, low cost, storage and Computing power. This has led to the daily production and accumulation of data of immense complexity in terms of volume, speed and variety which often hide useful information, making the situation pathetic [15, 16].

As information is one of the most important and expensive resources in the contemporary world in which we live, the massive collections of data collected and stored daily in Cloud servers have created the need to analyze it intelligently and efficiently in order to transform it into relevant information and knowledge that supports decision making. In this context, Data Mining has emerged as an effective tool for data analysis and the discovery of relevant information.

Data Mining has emerged as one of the most spectacular and challenging areas of our time, encompassing a variety of

Manuscript received July 24, 2020; accepted December 16, 2020.
*Corresponding author: abirachache@gmail.com

techniques and tools to meet the aforementioned requirements. As an interdisciplinary sub-domain of computer science, its birth has not occurred by coincidence but rather from the effective consolidation of various domains that have matured over time such as: machine learning, artificial intelligence, statistics, data and database management, information retrieval and visualization. Although Data Mining concepts were introduced a long time ago, the term "Data Mining" was widely adopted and accepted in the middle of the 1990s. According to the Gartner Group [17] and the authors of [16, 18]: *"Data Mining is the process of analyzing and discovering significant correlations, models, rules and trends in a large amount of data stored in repositories, using automatic or semi-automatic means, model recognition technologies, and statistical and mathematical techniques"*. Data Mining is the process of advanced analysis and non-trivial extraction aimed to find hidden patterns and autonomous implicit information, to discover correlations and rules from huge datasets previously collected; in order to interpret them into potentially useful knowledge and form computer models capable of adapting to particular situations. The process of data mining known as "knowledge discovery in a database" (KDD) is an iterative process that people often confuse with the term "Data Mining" and treat them as synonyms, knowing that Data Mining is considered as a step in the iterative KDD process (Fig. 1). This latter is founded on the fusion of the seven successive key steps [4, 10, 15, 19, 20, 21]:

Fig. 1. Knowledge discovery in databases process.

Today, Data Mining has given rise to a wide variety of different design learning strategies and techniques that can be grouped into two fundamental classes as shown in the Fig. 2 according to the objective of the exploration process, namely 1) *prediction* allowing the development of models based on current or previous (historical) knowledge in order to project the future state before its production [16, 19, 22, 23], and 2) *description* allowing the development of discovery and analysis models with useful syntheses based on the inspection of data and the description of their features [15, 16, 22].

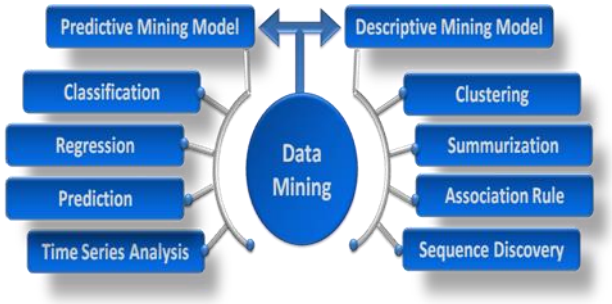

Fig. 2. Data mining process goal and strategies.

In this way, the incorporation of data mining tools and techniques can bring positive gains in many areas involving large amounts of integral and highly sensitive data by focusing them on the most precious information. In consequence, the integration of Data Mining techniques into Cloud Computing has become very important. Data Mining in the Cloud is the process of extracting structured information from unstructured or semi-structured web data sources. It will allow users to extract information and knowledge useful for decision-making from a virtually integrated database, which will contribute to reducing personal storage and infrastructure costs and provide efficient and secure services to their users by improving their quality [4, 10]. It can also help them learn lessons, optimize their information searches and uses, make decisions, reduce costs, and make future projections by transforming rigid data into useful data. It has become obvious that the future is highly expected to witness the real power of large amounts of data and its efficient treatment on Cloud Computing platforms through the use of Data Mining.

The remainder of this paper is organized as follows: Sections 2 explains how Data Mining and Cloud Computing, particularly the SaaS model, can be used in a reciprocal way. Important research for Data Mining in Cloud Computing, especially in the SaaS model, was presented in Section 3. Section 4 concludes the paper with future directions for the research.

II. Data Mining and Cloud Computing

The combination of Cloud Computing and Data Mining has an enormous advantage and potential. It can provide solutions to overcome some of the problems posed by both paradigms.

1. Cloud Computing Based On Data Mining

Cloud Computing is the new trend in Internet services that rely on remote server Clouds to manage tasks, and is a key area on which Data Mining must be concentrated. Data Mining in the Cloud is the process of extracting structured information from unstructured or semi-structured web data sources. Data Mining in the Cloud allows users to retrieve useful information from a virtually integrated data warehouse to make effective decisions and predict future trends and behaviors, which reduces infrastructure and storage costs. With this integration, the desired information on the behaviour, habits, interests and geographical location of customers can be found with a few mouseclicks.

Cloud Computing refers to the provision of computing resources over the Internet. Instead of data storage, maintenance and updating of hardware and software being

done locally by users, they will be provided by Cloud servers as a service to these users via the Internet. Data Mining in Cloud Computing allows organizations to centralize software and data storage management, ensuring that Cloud providers can deliver cost-effective, efficient, reliable and secure services to their end users [4, 10, 18, 24, 25].

Cloud Computing is an advanced model of distributed Computing, the integration and development of automatic techniques for extracting models and knowledge from the Cloud provides an intelligent Cloud by improving the quality of the services delivered to users. Companies and customers use these services to obtain information on different topics and take measures based on the presented data. The studies and research carried out have shown that Data Mining algorithms play a vital role in Cloud Computing environment and offers considerable potential for analysis and extraction of useful information and knowledge in different areas of human activity: commerce, networks, economics, medicine, biology, pharmacy, advertising, etc [23, 13].

✓ *Data Mining In The SaaS Cloud Model:*

The use of Data Mining is the application of techniques such as association rules, machine learning, clustering... etc. on a raw data set for exploration purposes. The use of Data Mining in a SaaS Cloud rivals the technique that could predict user behavior and learn their profile each time they interact with a service to provide a better personalized experience. The output data is visualized in terms of interactivity with the application/software. Usage data can be represented by relational tables or graphs that will then be used by companies or application developers to make informed decisions and predict future trends and behaviors. The integration of Data Mining techniques and applications is very necessary in the SaaS model of Cloud Computing. As SaaS services are provided via the Internet and user data is stored on the provider's site, the provider must be able to ensure the provision of flexible, secure and highly available services to satisfy the needs of its users. These needs are usually negotiated through a service level agreement between the customer and the service provider to ensure the performance of Cloud services.

Data Mining algorithms enable the quality assessment of potential SaaS software services on Cloud Computing by building different exploration models based on customer requirements. These algorithms are very useful for software service providers to evaluate their own services in order to solve some of the notable problems in SaaS services and to increase their availability and performance in the Cloud environment, based on the demands and needs of Cloud users. They also help Cloud users to be more flexible, to access to the technology rapidly and to evaluate their own satisfaction with the potential software services available in the Cloud Computing environment. However, the integration of Data Mining into SaaS Cloud Computing services should be robust enough to support the variation and increase in the amount of data produced and contribute to the efficient use of this data [15, 12, 17, 25].

✓ *Designing a SaaS Data Mining Process :*

A SaaS Data Mining process, in Fig. 3, can be modeled through a set of successive steps, these steps must be preceded by pre-processing steps:

Fig. 3. SaaS data mining process steps.

1) *SaaS Data Extraction:* Generally, the required SaaS data for such processing is distributed everywhere and is not stored in any type of data repository. As a result, developers must effectively manage and analyze huge amounts of distributed data, which presents a challenge for them. Thus, they must be able to extract and collect only the necessary source SaaS data and separate it from the existing source data.
2) *SaaS Data Transformation:* This step is based on the use of appropriate methods to treat inconsistencies and noise from the collected SaaS data and to transform them into a standard format. After the extraction phase of the SaaS data, a cleaning is applied to these data to get rid of noise and irrelevant data. Since we are in a dynamic and scalable environment, SaaS data flows are of massive volume and in rapid and continuous evolution, which involves a lot of difficulty during their storage in data warehouses. For this purpose, it is essential to find interesting models among the raw data by performing a multidimensional analysis on aggregate measures such as sum and mean. Then, the resulting data must be assembled in appropriate standard formats.
3) *Application of Data Mining Algorithms:* During this phase, SaaS data stored in multidimensional formats is loaded for use as input to various algorithms and Data Mining methods. According to the predefined needs of the desired model, a descriptive and/or predictive Data Mining technique (s) is (are) selected and subsequently applied to the SaaS data. After application of the Data Mining algorithm, the output will be generated in terms of performance of SaaS services provided on the Cloud by service providers.
4) *Obtaining Exploration Results:* Finally, the results obtained can be used by service providers to analyze the developed model and determine which Data Mining method is most effective and best adapted to the given SaaS data in terms of response performance. On the other hand, they will help service users to make a decision that is appropriate to their software service needs.

2. *Cloud-Based Data Mining*

The rapid development of information technology and storage capacity has led to the accumulation of a huge amount of diversified data. In a reciprocal need to what was discussed in the previous subsection, the massive processing of this data

has become an important problem in the field of Data Mining. In the past, the Data Mining process was carried out on a high-performance machine or large computer equipment, which became inefficient. Therefore, the question to be resolved is how to improve the parallelism and efficiency of Data Mining algorithms. In such situations, Cloud Computing is a powerful support and an effective way to overcome Data Mining problems through its enormous capacity to: process large amounts of data, storage, computation and elastic changes. By including Cloud Computing in Data Mining, major data processing and storage problems, ownership costs of Data Mining tools and techniques will be solved.

Cloud Computing adopts the same principle of providing software and hardware to provide a Data Mining task where the latter is provided as a service on the Internet. Cloud Computing-based Data Mining is a service-oriented, approach based on the "Cloud", providing an interface for Data Mining services to a variety of users. Users can use the various services via the interface provided by the system without any need for pre-knowledge of the system's operation and concern about the system's computing and storage capacity. They simply need to select the appropriate algorithm to process their data and run it to obtain the results of the Data Mining. In such a Cloud-based approach, the Cloud-based Data Mining system can adopt a pay-per-use method that allows companies or individuals to obtain the desired service directly and get rid of software purchase expenses and the obstacles that prevent them from taking advantage of Data Mining techniques [9, 24, 26, 27].

Cloud-based Data Mining services can address a particular process in the Data Mining process, a Data Mining technique, distributed Data Mining models and a KDD process. The authors in [28], summarized four levels of Data Mining service in Cloud Computing as shown in the Fig. 4:

Fig. 4. DMC service levels.

1) *Single KDD Steps*: At this level all the steps included in the KDD process including: selection, representation and visualization are expressed as services.
2) *Single Data Mining Tasks*: It includes separate Data Mining services such as: prediction, classification, clustering and discovery of association rules.
3) *Distributed Data Mining patterns*: at this level distributed Data Mining methods are implemented such as collective learning, parallel classification as a service.
4) *KDD Process*: This level includes previous tasks and models composed in a multi-step work process.

This conception is incremental where the implementation of services in a level is based on the services of the levels preceding it, which allows for the reuse of processes, techniques and models already available. This framework can be used to develop distributed Data Mining tools as a composition of unique services that can be accessed at any time and from anywhere [28].

In terms of the cost of using Data Mining techniques, integrating the Cloud into Data Mining offers an advantage for small businesses by allowing them to have the possibility of renting a Data Mining service in the Cloud with a minimal cost for an efficient analysis of all the organization's data that was previously reserved only for large companies [27].

✓ *SaaS Cloud-based Data Mining:*

The Software as a Service "SaaS" model is considered as the king of all Cloud services. It reduces costs by offering flexible licensing options and externalizing hardware effort. In SaaS, services are centrally managed and remotely accessible on the Web. In other words, no software components are installed at the customer's site, they are located on the server of a software service provider that handles hardware, software updates and technical maintenance. The provision of SaaS services is maintained according to a multi-tenant architecture that involves a single physical instance with customers hosted in separate logical spaces and according to an architecture where there can be multiple variations in how a single instance is actually implemented and how multi-tenant is actually achieved [18].

As SaaS refers to computer software provided as Internet services, in the SaaS model of Cloud Computing, Data Mining software is also provided in this way. The integration of the SaaS model into Data Mining allows providers to define software capable of processing and providing a Data Mining task to users as an on-demand service over the Internet. This is reflected positively on customers in terms of reducing the costs of using Data Mining tools and techniques. On one side, the customer only pays for the Data Mining tools he needs, which allows him to save a lot compared to the complex Data Mining suites he doesn't use in an exhaustive way. On the other side, he only pays for the costs generated by the use of the Cloud. He doesn't need to maintain a hardware infrastructure; he can apply Data Mining simply via his browser. This reduces the obstacles that prevent small businesses from benefiting from Data Mining [18, 25].

III. Related Works

Currently, the coupling of Data Mining algorithms with Cloud Computing technology has become the research trend. This is reflected in the amount of scientific work done in the field to overcome some of the problems posed by both paradigms. While Data Mining has been widely used in Cloud Computing, the research that addresses its integration into the SaaS Cloud Computing model remains limited. In this section of the article, various important research studies in the field of Data Mining in Cloud Computing were presented. We have divided them into three main categories: approaches for clustering in the Cloud Computing paradigm, approaches for Classification in the Cloud Computing paradigm and approaches for Data Mining in the SaaS Cloud Computing model.

1. Clustering-Based Approaches

In the context of Cloud Computing, performance improvement must be maintained by ensuring: efficient scheduling and execution of application tasks, fast and effective access and recovery of resources, and load balancing. Different solutions are proposed to improve the quality of Cloud services on one side and to meet the users' requirements on the other side.

Patki in [29] studied the clustering algorithms that can be applied with high precision in Cloud Computing, divided into two categories including Hard Clustering algorithms (K-means, hierarchical clustering) in which each data belongs to a single group and Soft Clustering algorithms (fuzzy C-means) where one data can be part of several groups. The author noticed that Soft Clustering techniques are more efficient and preferable in Cloud Computing environments as Cloud data are heterogeneous in nature and may contain similarities and have relationships with several groups making their retrieval fast and accurate. The result of this discussion was proven by Aparajita et al. in [30] who provided a comparative implementation of a set of clustering techniques (based on partitioning, based on hierarchy) on data in Cloud Computing to assess the performance of each one on the load balancing issue.

Madhuri and al. in [13] introduced an improved extension of the Agglomerative Classification algorithm for Data Mining in the Cloud. In the proposed clustering model, the data are initially considered as a single cluster. The similarity and dissimilarity parameters are calculated between each pair of objects and are merged into a group based on their degree of similarity or considered noise based on their degree of dissimilarity. A binary hierarchical cluster tree is used to group data entities on the Cloud. In more enhanced parallel work, Madhuri et al. in [31] proposed a Cloud data clustering approach that applies the improved Hierarchical Agglomerative Clustering algorithms "CURE" over a heterogeneous network. The proposed approach uses MapReduce to parallelize the system and divide the data into subparts in order that recovery from Cloud storage can be used, which increases efficiency and improves system latency. The experimental results show the efficiency of the proposed algorithm in terms of reduction of execution time, scalability and availability. In a similar approach, Srivastava et al. in [32] proposed the implementation of the Hierarchical Agglomerative Clustering algorithm adapted to large data in order to increase the efficiency of the algorithm by performing tasks in parallel. The basic idea is to read the subsets of the database, apply the clustering algorithm, combine the results with those of the previous samples and proceed in this way until all the data is available in the main cluster. The results demonstrate the efficiency of this algorithm which increases with parallelism on the Cloud-based architecture.

Sarkar et al. in [36] proposed an approach that uses a Hierarchical Clustering algorithm to organize data according to the type of data stored in Cloud data centers. The algorithm allows to form clusters of data by classifying them according to their nature/source to be easily retrieved by end users. This approach provides significant performance on large heterogeneous datasets allowing fast and efficient data access.

Atan in [35] proposed a new approach based on Clustering algorithms (K-Means and Hierarchical Agglomerative Clustering) to ensure the accuracy of the adequacy of the user needs in the process of obtaining services from the service providers. The platform allows users to access to their service providers by obtaining a result that is easy to browse and quickly retrieve relevant information based on their interests. The proposed approach is able to increase the level of user satisfaction and obtain guaranteed services for their requirements.

Shindler et al. in [33] proposed a fast k-Means algorithm applicable to large data with sequential access. The algorithm uses the notion of the approximate search of the nearest neighbor to calculate the allocation of installations from each point in this approach. The proposed algorithm outperforms existing algorithms providing advanced performance in theory and practice proven by faster execution time and a better approximation factor. Mahendiran et al. in [34] proposed the implementation of the K-Means Clustering algorithm in a Cloud environment. The algorithm was implemented in Google Cloud using Google App Engine with Cloud SQL. The experimental results prove the performance of the algorithm in the Cloud on the popular IRIS data set.

Asnani in [37] proposed a Clustering model of huge text data based on feelings using the K-Means algorithm to find the hidden feelings of text editors. The technique is based on classifying the text according to its orientation in terms of the user's mood in the available text communications. The model architecture uses Haddop for the storage of the data to be processed after their pre-processing. In addition, the data is used with the API of the NLP tool that is used to produce the data tag. Then, the marked data is retrieved and used with the improved k-means clustering. The proposed algorithm was developed in JAVA to be deployed as a Cloud service and evaluated on different performance parameters. It has proven its efficiency with high accuracy, low error rate and optimization in terms of time and memory consumption during analysis.

Panchal et al. in [38] proposed a dynamic algorithm for allocating virtual machines using Clustering to improve performance and maintain load balancing in the Cloud Computing environment. To plan virtual machines, the K-means algorithm is used to form clusters. The evaluation results of the proposed algorithm by CloudSim show its performance in Cloud data centers in terms of load balancing and efficient processor utilization.

Comparison of clustering algorithms implemented in cloud computing is shown in Table I.

TABLE I. COMPARISON OF CLUSTERING ALGORITHMS IMPLEMENTED IN CLOUD COMPUTING

Ref N°	Used Algorithms	Used Framework	Type of Managed Data	Objectifs	Advantages	Main Considered Features									
						Parallelism	Scalability	Load Balancy	Efficiency	Availability	Fault Tolerance	Robustness	Execution Time	High Dimensionality Handling	Noise Handling	Accuracy
[13]	improved hierarchical agglomerative clustering	Cloud Environment	Un-Structured	Efficient retrieval of a huge amount of heterogeneous data in the Cloud Computing.	Improved efficiency and performance in exploitation of massive amounts of noised data. Optimization of the number of clusters.	Not used	✓		✓				✓	✓	✓	✓
[]31	Improved Agglomerative CURE Hierarchical Algorithm	Map Reduce, Haddop	Structured	Efficient extraction of a huge amount of noisy data over a heterogeneous network.	High efficiency, scalability and robustness to manage large datasets. Reduction of the execution time.	Input Data	✓		✓	✓	✓	✓	✓	✓	✓	✓
[32]	Hierarchical Agglomerative Clustering, Virtual K-Means	Cloud Environment	Structured	Parallelism of data and improvement of the efficiency and adaptability of clustering large databases.	Optimisation of algorithm efficiency and execution time via the implementation of macro and micro clustering on large and small datasets.	Data	✓		✓				✓			✓
[33]	Hierarchical Clustering	Cloud Storage Environment	Un-Structured	Efficient organization of massive data volumes in the Cloud	Scalability and efficient retrieval of massive amounts of data, agility and fast access, provide statistics of efficient use of cloud storage space.	Not used	✓		✓				✓			✓
[34]	K-Means, Agglomerative Hierarchical Clustering	Cloud Framework	Un-Structured	Satisfy users' need for quality of service in a cloud environment with large amounts of data	Reduction of inaccuracy of requirements and increase in customer satisfaction. Economy of search and access time to services. Easy, fast and efficient exploration.	Not used	✓						✓			✓
[35]	K-Means Clustering	Cloud Framework	Structured	Improved k-means for optimal management of large streaming data	Minimized k-point selection cost, faster execution time and better approximation factor, gains accuracy.	Not used	✓		✓				✓			✓
[36]	K-Means	Google Cloud using Google App Engine.	Structured, Un-Structured	Optimized management of large volumes of data and real-time data	High efficiency in the analysis and storage of large databases. High clustering accuracy, cost reduction and profile maximization.	Not used	✓	✓			✓		✓			✓
[37]	Distributed K-Means	Hadoop Storage, HDFC used in Cloud Environment	Un-Structured	Provide Clustering as a service and maintain cluster analysis performance of large datasets stored in the Cloud.	Handling High dimensionality and noisy data. Medium scalability and high accuracy. Reduced error rate and analysis time.	Data	✓						✓	✓	✓	✓
[38]	K-Means	CloudSim Environment	✕	Dynamic allocation of virtual machines and performance improvement in Cloud.	Improved scalability, fault tolerance and data availability. Load sharing and efficient CPU use.	Not used	✓	✓			✓		✓			✓
[39]	K-Means, Genetic algorithms, Particle Swarm Optimization	CloudSim, Cloud Environment	✕	Load balancing in a Cloud environment	Reduced task execution time and overall cost, Optimized speed, efficiency and performance, Improved flexibility, scalability and availability, Better load balancing and faster convergence.	Not used	✓	✓	✓	✓			✓	✓		✓
[40]	k-means; particle swarm optimization	Cloudsim Framework	✕	Optimization of the makespan for efficient scheduling tasks in a Cloud environment.	Improved of virtual machines use and load balancing. Simplicity, speed, efficiency and minimisation of complex task execution time.	Not used	✓	✓	✓				✓	✓		✓
[41]	K-Means	MapReduce, Mahout, Hadoop plateform	Un-Structured	Analyze and enhance performance gains with large datasets.	High Scalability and performance gains. Reduced overhead and execution time, Increased network utilization and reduced CPU utilization.	Data	✓	✓	✓				✓	✓		✓
[42]	K-Means	MapReduce, HDFS, Cloud Environment,	Un-structured	Improve the performance of massive data mining.	Enhanced large-scale data processing capability and performance, Reduction in execution time.	Data	✓					✓	✓			✓
[43]	Optimized parallel K-Means	Map reduce	Un-structured, semi-structured	Maintain high performance in large data mining	High efficiency, robustness and scalability. Reduced I/O and network costs. Improved cost and time of algorithm computation.	Data	✓		✓		✓		✓			✓
[44]	Web Fuzzy C-Means	Cloud Environment	Structured, Semi & Un-Structured	Optimization of data processing in the Cloud.	Scalable and efficient parallel processing, Enhanced system with ease and dynamics resources providing.	Data	✓		✓			✓	✓			✓

By sharing the same objectives, Sajjan et al. in [39] proposed an approach similar to the preceding one with the difference that the clusters formed by the K-Means algorithm are combined with the functionalities of three heuristic algorithms: genetic algorithm, simulated annealing and particle swarm optimization, which reduces the time required to find an appropriate virtual machine, thus reducing task duration and improving Cloud system performance.

Recently, Raju et al. in [40] proposed a hybrid approach called KPSOW which aims to associate the K-Means Clustering technique and the PSO bio-inspired optimization algorithm with the inclusion of the weight concept to perform effective task scheduling in a Cloud environment. The basic idea of the proposed work is to use the K-Means algorithm to separate Cloud tasks into groups of low complexity and high complexity tasks and calculate their weights. Then, assign low complexity tasks to low performance IT resources and high complexity tasks to high performance IT resources by minimizing lifetime using the PSO algorithm. KPSOW proves its simplicity in the fast execution of complex tasks, efficient use of IT resources and appropriate load balancing of virtual machines compared to FCFS and PSO methodologies.

In [41], the authors presented a framework for analyzing the performance of the Mahout framework with K-Means on large datasets running on Amazon EC2 instances. In this work, the vectors are converted into a specific Hadoop file format which is SequenceFile to be used as input for the MapReduce Framework. Vector processing with Mahout_K-Means generates centroid coordinates and samples assigned to each cluster. Experimental results produce better performance gains by reducing clustering time and CPU usage. In a parallel implementation of K-Means, the authors in [42], presented a K-Means algorithm based on the MapReduce framework. The proposed algorithm allows the division of data sets that will be used as inputs for MapReduce. At the level of each map, centers are selected and data are organized in clusters. Then, the outputs of all map functions are combined and reduced to form the k clusters and data points. Experimental results prove the superiority of the adaptability of the proposed algorithm over the traditional K-Means algorithm to efficiently manage massive data. In [43], the authors proposed an optimized and parallel K-Means algorithm based on the efficient MapReduce model to achieve high performance of large-scale data clustering. Using the probability sampling technique, the proposed algorithm estimates iterations so that it uses only certain subsets of the original large data sets. It starts by generating a sample from the original data set. After sampling, the mappers cluster the data using K-Means and obtain centres that will be merged into a reducer to produce k final centres using two new merging methods introduced by the authors: weight-based merging clustering (WMC) and distribution-based merging clustering (DMC). Then, these k centers are used to generate the Voronoi diagram to partition the original data set and obtain the final clustering result. The experimental results show the superiority of the proposed algorithm in terms of efficiency, scalability and robustness.

An efficient mining algorithm based on Web Fuzzy Clustering analysis of large data sets in Cloud Computing has been presented by Xianfeng Yang et al in [44]. The algorithm uses a fuzzy web object similarity matrix and provides services to users via an interface. It starts with data cleansing using call services. Subsequently, the data layer deposited in the Cloud Computing platform provides storage space for data mining services. The FCM algorithm can be used for Fuzzy Clustering. Experimental results show that this method can effectively improve data mining performance.

2. Classification-Based Approaches

Cloud Computing provides computing resources via the Internet with elasticity according to a pay-per-use model. Efficient task scheduling and parallel processing are the key to maintaining better resource utilization and load balancing and to satisfying scalability and performance requirements in the Cloud Computing environment. As Cloud services are delivered over the Internet; preserving data security and confidentiality of Cloud resources and services are the major preoccupations for Cloud providers as well as for Cloud users. Cloud systems must be able to protect the privacy of users of Cloud services from attacks and malicious behavior by network nodes that can seriously affect network capacity and performance. In recent years, research and application of classification algorithms in Cloud Computing have been rapidly developed and widely used. These algorithms make a remarkable impact on the performance of the Cloud Computing environment through its ability to solve complex problems as mentioned above.

Arjmand and al. in [45] proposed a fuzzy KNN classifier to classify Cloud data according to their privacy levels and apply different encryption methods to secure the data before transferring it for storage in the Cloud. The algorithm allows data to be grouped into three classes: confidential, private and public in order to encrypt confidential and private data very carefully using successively the RSA and AES encryption algorithms, which conserves time and system resources. Then, the classified data is sent to the Cloud for storage. Wang in [46] introduced a new approach that addresses the issue of Data Mining in the Cloud while preserving confidentiality. The author proposed an algorithm called PPKC (Privacy Preserving K-NN Classification) based on K-Nearesr Neighbor to extract and classify data in the Cloud while avoiding the revelation of users' private and sensitive information. To ensure the accuracy and efficiency, the proposed algorithm uses the BWC (Binary Weighted Cosine) metric to measure the similarity of records during which the private correspondence protocol is used to protect data confidentiality. Bajare et al. in [47] proposed a secure k-NN classification technique that preserves the confidentiality of encrypted data in the Cloud using the semi-honest model. The proposed technique uses the Elliptic Curve Cryptography (ECC) encryption algorithm to protect data confidentiality, user input requests, task results and hide data access patterns. In an analogous context, Goutham .V and all in [48] proposed a protocol similar to what was presented previously with the absence of semi-honest models and the use of Standard Homomorphic Encryption methods to encrypt data. These models provide an important data security advantage. Kour in [49] proposed a classification technique to secure data in a Cloud Computing environment. The proposed technique uses an improved Bagging and Boosting algorithm to classify data

into sensitive (private) and non-sensitive (public) data. Then, the Blowfish algorithm is used to secure sensitive data while non-sensitive data is sent to the Cloud without encryption. In order to protect data availability in case of a malicious attack, partitioning is applied. CloudSim simulation results show that an improved bagging technique gives better results than the K-NN classification algorithm, thus reducing classification time and improving accuracy.

In [50], the authors proposed a network intrusion detection system (NIDS) based on the Naive Bayes classification method combined with Snort to detect and prevent malicious activity and network attacks in an IaaS Cloud. NIDS uses Snort in the first place to perform signature-based detection that can detect known attacks. Then NIDS uses the Naive Bayes classifier to perform anomaly detection, which detects unknown attacks and determines if a given behavior is malicious or not by observing previously stored network events. The proposed system guarantees a high accuracy and a reduced detection time with an affordable calculation cost. Ebadifard and al. in [51] provided an algorithm for dynamic task planning in a Cloud environment based on the Naive Bayes classification method to maintain load balancing of virtual machines. This algorithm allows to classify virtual machines and to select in a balanced way a machine adapted to existing requests by transferring requests from an overload machine to under-loaded machines, which reduces makespan time and increases the load balancing level. Zhou et al. in [52] proposed a fast parallel classification algorithm in the Cloud environment. The algorithm consists of the probabilistic model Naïve Bayes based on MapReduce. The proposed algorithm includes a training step using the Map and Reduce politic and a prediction step to predict the data recording with the output of the training model. The experimental results show that the proposed algorithm can not only effectively handle large datasets but also improves the efficiency and performance of the original algorithm. Another approach presented by Qing He and all in [53] recommends a parallel implementation of classification algorithms, including k-nearest neighbor, Bayesian model, decision tree based on MapReduce. This has allowed the resulting parallel algorithms not only to be applicable to the exploitation of large data sets but also to have the required property of linear scalability.

Kamdar et al. in [54] proposed an approach based on Naive Bayes and Support Vector Machine (SVM) classification algorithms for analyzing and classifying large data in Cloud Computing. The proposed algorithm initializes the weight of the learning data and then creates a new data set using selection with the alternative technique. After that, it calculates the a priori and conditional probabilities of the new dataset, and classifies the learning data with these probability values. The weights of the classified examples are updated according to their classification accuracy. The process of creating data sets is continuous until all learning data is correctly classified. To classify the new data, classifier votes are used. The output of this step is provided to SVM for further classification. In this approach both algorithms are used with MapReduce providing high accuracy, efficiency and performance.

In [55], Catak et al. proposed an approach called the SVM Cloud Training Mechanism (CloudSVM) which consists in implementing a distributed SVM based on the MapReduce technique to improve scalability and parallelism in Cloud Computing systems. The basic idea of the proposed approach is that first, the SVM algorithm is trained on distributed Cloud storage servers that run in parallel. Then, the support vector in each formed Cloud node is collected and merged. These two steps are repeated until the SVM converges on the optimal solution. The experimental results proved the efficiency of the proposed approach in the Cloud Computing environment. A similar approach has been proposed by Lu Shuhong in [56] unlike the introduction of the penalty factor to manage the dynamic of Cloud data and improve the accuracy of Data Mining algorithms in Cloud Computing. The proposed algorithm presented higher accuracy, improved exploration time and performance.

Kumar et al. in [57] discussed the three scheduling techniques Min-Min, Max-Min and genetic algorithm as well as the performance analysis of Min-Min and Min-Max was shown. An improved genetic algorithm has been proposed based on the combination of the two algorithms Min-Min and Max-Min for scheduling tasks in Cloud Computing. The basic idea of the proposed algorithm is to generate the initial population of the genetic algorithm using Min-Min and Max-Min, which can provide a better initial population than if we choose the initial population at random resulting an improved genetic algorithm. The proposed algorithm provides high performance, better resource utilization and reduced global execution time for tasks compared to standard genetic algorithms. Zhu et al in [58] proposed a hybrid approach that combines GA and multi-agent techniques for designing a multi-agent genetic algorithm (MAGA) as a way to achieve load balancing between virtual machines during the scheduling of tasks in the Cloud Computing environment. The proposed algorithm treats a person within GA as an agent capable of local perception, competition, cooperation, self-learning to achieve the objective of improving the quality of global optimization results compared to standard GA through the interaction between the agent and the environment. The results show the efficiency and superiority of the proposed algorithm compared to Min-Min and prove that MAGA is able to achieve better CPU utilization and load balancing performance. Le et al. in [59] introduced a non-dominated genetic sorting algorithm based on grid partitioning (NSGA-G) to improve the diversity, efficiency and convergence characteristic of current non-dominated genetic sorting algorithms "NSGA". The proposed algorithm uses partitioning to divide the solution space into several small groups so that it calculates and compares only the solutions in a group. In order to maintain diversity, the partitioning strategy is introduced in the random selection of the solution by choosing a random group, which reduces the calculation time. The proposed algorithm achieves better performance and better quality and computation time than other algorithms, such as NSGA-II, NSGA-III and MOEA / D.

Ref N°	Used Algorithms	Used Framework	Type of Managed Data	Objectives	Advantages	Limitations	Main Considered Features							
							Parallelism	Scalability	Load Balancy	Security	Efficiency	Resource Optimization	Execution Time	Accuracy
[45]	Fuzzy K Nearest Neighbors Classifier	Cloud Storage	Structured	Data security in the Cloud	High efficiency. Very careful encryption of confidential data, Reduction of the total data security time, Improvement of the recognition rate. Time and system resources economy.	Classification execution time is slightly higher	✗	✓		✓	✓	✓	✓	✓
[46]	K Nearest Neighbors (KNN)	Cloud Framework	Structured , semi-Structured	Preserving the confidentiality of Cloud data during its exploration	Reduced error rate and improved security, High efficiency and accuracy, medium Scalability.	Declining performance with many explicative variables.	Data	✓		✓	✓			✓
[47]	k Nearest Neighbors	Cloud Model	Structured	Maintain the security and confidentiality of user data in the Cloud.	High Security, Scalability and Efficiency	less efficient solution to the SMIN n (Secure Minimum) problem	✗	✓		✓	✓			
[48]	K Nearest Neighbors	Cloud Storage	Structured	Preservation of confidentiality in data mining for the encrypted database stored in the Cloud	High security, medium scalability Efficient classification of encrypted data.	Low efficiency of the SMINn protocol.	✗	✓		✓				✓
[49]	Bagging, Boosting	Cloud Storage	Structured	Data security for the Cloud environment	Good data security with time and cost economy.	Limited improvement in accuracy and time of classification	✗	✓		✓			✓	✓
[50]	Bayesian Classifier, Snort rule.	Cloud Environment	Structured	Detection and prevention of intrusions and attacks in the Cloud	Higher rate and shorter time to detect intrusions, reasonable calculation cost. high efficiency, scalability and accuracy,	Optimal positioning of the Network Intrusion Detection System.	✗	✓			✓		✓	✓
[51]	Naive Bayes Classifier	Cloud Computing	Structured	Optimization and dynamic scheduling of tasks in the Cloud.	Good improvement of Makespan and load balancing degree of VMs, Higher accuracy and speed, increases efficiency and resource utilization.	Disregards the criterion of cost reduction for service providers	✗	✓	✓		✓	✓	✓	✓
[52]	Naive Bayes Classification	Map-Reduce, Haddop with HDFS	Structured	High efficiency and scalability in Big Data Classification	Improved latency and response time, High Scalability and good parallelism	Unoptimized exploitation of IT resources	Input data	✓			✓	✓		✓
[53]	K Nearest Neighbors, Naive Bayes, Decision Tree	Map-Reduce Hadoop	Structured	Maintain high performance and accuracy in Big Data classification	High promising scalability due to the good parallelism.	A low-level of efficiency in the use of IT resources.	input dataset	✓			✓	✓		✓
[54]	Naive Bayes and SVM	Hadoop, Map-Reduce	Structured	Fast and efficient analysis of large amounts of Cloud Computing data.	High accuracy, scalability and reliability, improved efficiency and performance.	Long processing time and less efficiency for small data sizes	Data	✓			✓		✓	✓
[55]	Support Vector Machine (SVM)	Cloud Storage Hadoop, Map Reduce	Structured , un-structured	Maintain Efficiency in Large Amount of Training Data Mining	Excellent precision and stability, improved generalization, High of scalability and parallelism.	parameter sensitivity	Training data	✓					✓	✓
[56]	Improved SVM	Cloud Environment, Map-Reduce	Structured	Efficient Cloud Computing Data Mining	High Classification Accuracy, Improved information mining time and Medium Scalability.	Moderate execution time compared to existing algorithms.	Data	✓					✓	✓
[57]	Improved Genetic Algorithm	Cloud Environment	Structured	Efficient task planning and better use of resources	Minimisation of makespan. Efficient use of resources. Improved performance and High Scalability.	Problems with local optimums	✗	✓			✓	✓	✓	
[58]	Multi-agent genetic algorithm (MAGA)	Cloud Environment	Structured	load balancing based on the management of virtualized resources in the Cloud Computing	Better CPU utilization and memory load balancing, reduced failure rate, high efficiency and optimization	High performance is sensitive to weighting factors.	✗	✓	✓		✓	✓	✓	✓
[59]	multi-objective genetic algorithm NSGA-G	Cloud Environment	Structured	Multi-objective query optimization to define data configurations and query processing strategy in Cloud.	Higher performance enhanced quality and computing time, improved diversity, efficiency and convergence.	Problems of approximate optimal solutions.	✗	✓			✓		✓	✓
[60]	SPRINT based on Tree Decision Algorithm	MAP-Reduce avec Hadoop	Structured	High level of parallelism in Cloud Computing environment mining	High efficiency and medium scalability, reduced execution time and improved performance.	Complexity of finding good split points for data partitioning during tree growth	Data	✓			✓		✓	
[61]	An Enhanced Very Fast Decision Tree Algorithm	Cloud environment	Structured	Detecting effectively the occurrence of Distributed Denial of Service attacks.	High accuracy for detecting and classifying attacks of streaming data in real time, reduced resource consummation (time/memory), efficient handling of noisy data.	Robustness failures due to defects in the decision tree classifier. Low Scalability.	Data	✓		✓		✓	✓	✓

TABLE II. COMPARISON OF CLASSIFICATION ALGORITHMS IMPLEMENTED IN CLOUD COMPUTING

Zhang et al. [60] proposed a classification strategy to extract Cloud data using an improved "SPRINT" classification algorithm based on the decision tree to adapt to the large-scale Cloud environment. The proposed algorithm has been implemented above the MapReduce framework to ensure parallelism of Data Mining tasks that can be performed on multiple nodes. The main concept of the proposed algorithm is to divide the learning data and submit them to different calculation nodes that are responsible for processing the attribute list and recursive data partitioning. The partitioning process is carried on until the members of each partition are very similar or the size of the partition is very small. Latif et al. in [61] proposed an Enhanced Very Fast Decision Tree (EVFDT) algorithm that can effectively detect occurrences of distributed denial of service (DDoS) attacks and classify them in a Cloud-assisted WBAN (Wireless Body Assist Network) environment. The architecture of the proposed Distributed Denial of Service (DDoS) attack detection system studies the behavior of network traffic to form a learning database for the decision tree. New incoming traffic is classified as attack or non-attack by building a classification tree based on the EVFDT. If a DDoS attacks is detected, an appropriate tracking mechanism is applied to track an attacker and block his traffic. The EVFDT algorithm has proved its ability to detect attacks with significantly high classification accuracy and a low false alarm rate with less memory overload. Comparison of classification algorithms implemented in cloud computing is shown in Table II.

3. Data Mining-based SaaS

As SaaS services emerge, evaluating and improving the quality of the SaaS services provided becomes more and more essential for both clients and Cloud service providers. Recently, several researchers have been exploring the use of different Data Mining techniques in the SaaS model to: 1) measure the goodness of SaaS solutions by taking into account the status of customer requirements to enable them to distinguish between Cloud service providers, 2) allow service providers to have some measure to know the superiority of Cloud services in order to improve their solutions.

Kanagalakshmi et al. in [62] presented a model for estimating the quality of SaaS software services in Cloud Computing based on Clustering Data Mining techniques. The proposed evaluation model includes steps to follow for SaaS data: 1) Extract SaaS data, 2) Transform SaaS data, 3) Load to RDBMS/MDBMS format, 4) Apply the Clustering algorithm, 5) Obtain the appropriate cluster. The proposed Cluster Model helps service providers to increase the availability and scalability of software services in the Cloud Computing environment. It also helps Cloud users to evaluate the potential software services available in the Cloud environment. For the similar objective, the same steps were followed by Dhanamma et al. in [12] although in step 4, the authors used the most popular hard hierarchical clustering algorithms.

Dhanamma et al. in [63] proposed an approach for SaaS evaluation based on the Constraint Based Clustering (CBC) algorithm. The proposed model takes into account customer requirements as well as the quality attributes of SaaS services. It uses k-Means clustering to build clusters that will be transformed into micro clusters according to user-specified constraints, allowing the identification of the impasse. If the impasse is found then the final cluster forms, otherwise "break the micro cluster" and repeat the operation until an impasse is found.

Dhanamma et al. in [64] provided various Data Mining clustering algorithms (K-Means Clustering, B-Cluster, Hierachical) to assess the quality of SaaS. The purpose of this research is to provide a decision making system with optimal solutions for users and Cloud service providers to quickly and easily select / provide potential software services according to specified requirements. Experimental results show the superiority of K-Means algorithm in terms of execution time compared to other methods. For maintaining the above objectives, the same authors in [65] used the Clustering "Partitioning Around Medoids" (Pam) algorithm. After the collection and processing of SaaS attribute data, a cluster analysis is called allowing the definition of the optimal number of clusters using the elbow method and the formation of the data clusters by applying the PAM algorithm. The experimental results show the efficiency and robustness of the PAM algorithm compared to K-Means in the presence of noise and aberrant values.

Kamalraj et al. in [66] proposed an approach based on Data Mining techniques for the analysis of the market basket of Cloud Computing products. The proposed approach relies on a Clustering algorithm to create groups of SaaS products and uses "Association Rule Learning" to find the relationships between the different data elements in the available clusters. In addition, it uses "Market Basket Analysis" to analyze services groups, identify services frequently used by customers and associates by new products as a "free trial" version. The proposed approach has mainly minimized the time required to find the required products in order to meet customer expectations and to make little effort to match them with new products to increase business gains.

Dhanamma et al. in [67] proposed a model to evaluate the quality of SaaS in the Cloud Computing environment based on the Estimation and Maximization (EM) clustering algorithm. First, the approach collects SaaS data on which pre-treatment techniques are applied. Then, it uses the EM-Clustering algorithm to form data clusters. The last step is to visualize the output data, which allows the attributes to be interpreted and evaluated. In a similar study [68], the same authors proposed a quality model called SAASQUAL (Software as a Service Quality model) to evaluate SaaS in the Cloud Computing environment and differentiate the quality of SaaS providers based not only on the EM-Clustering algorithm but also on different attributes and metrics that measure software quality. The Clustering-EM demonstrates its efficiency for incomplete data sets and proves its ability to help Cloud users select their best SaaS services and SaaS providers improve their products to build customer loyalty in the competitive world.

In [69], Yusoh et al. proposed a penalty-based Genetic Algorithm for the problem of SaaS composite Cloud placement, which considers not only the placement of SaaS software components, but also the placement of SaaS data. The problem can be classified as an optimization problem whose purpose is to optimize SaaS performance based on

its estimated execution time which depends on the time spent searching for the placement solution for each configuration. This is the first attempt at SaaS placement with its data on the Cloud Provider's servers. In [70], the authors developed a Genetic Clustering Algorithm (GGA) for clustering multiple composite SaaS into Cloud Computing clusters to address only the issue of composite SaaS placement and SaaS resource optimization. The proposed approach minimizes the use of SaaS resources without violating their SLAs by reconfiguring the location of application components while maintaining application performance, balancing thermal distribution between servers and minimizing data consumption. An algorithm that swings between the two latest works has been proposed in [71]. The proposed algorithm is similar to what was presented in [70] unlike the introduction of the notion of penalty used in the algorithm proposed in [69] knowing that it doesn't consider the placement of SaaS data. These researches demonstrate the feasibility and scalability of the GA. Comparison of datamining algorithms implemented in SaaS is shown in Table III.

TABLE III. COMPARISON OF DATAMINING ALGORITHMS IMPLEMENTED IN SAAS

Reference	Used Framework	Used Algorithms	Type of Managed Data	Objectifs	Advantages	Limitations	Evaluation Focused On			Quality Attributes					
							User perspective	Provider perspective	Quality Attributes	Eficiency	Resource Optimization	Noise Handling	Availbility	Scalability	Reusability
[12]	Cloud SaaS Environment	Various Clustering Methods	Un-Structured, Structured	Evaluation of SaaS on the Cloud.	Useful for users and providers to measure the quality of SaaS services.	Don't consider the software's quality attributes	✓	✓							
[63]		Constraint Based Clustering, K-Means	Structured	Multi-criteria evaluation of potential SaaS quality.	Increased service quality and reduced costs, greater customer satisfaction and loyalty.	Cluster quality metrics are less efficient.	✓	✓	✓	✓					
[64]		K-Means, Hierarchical, BiCluster	Structured	Optimization of decision making (selection/provision) on qualified SaaS services.	High speed and flexibility of the SaaS quality specification	Determination of algorithms input parameters require knowledge of the domain.	✓	✓							
[65]		Partitioning Around Medoids Algorithm	Structured	Evaluate the quality of any SaaS product with identified quality attributes in the Cloud.	High Performance and robustness in the presence of noise.	Weak analysis of large data sets, less stability.			✓	✓		✓	✓		✓
[66]		Association Rule Learning	Structured	Analyse the SaaS market basket and identify customer behaviour to increase business gains.	Reduced time and effort for advertising new SaaS products, high user satisfaction, increased financial returns	Low performance of specific algorithms for clustering and association.	✓	✓							
[67]		Expectation–Maximization algorithm	Structured	SaaS quality evaluation for selecting the best service according to users' needs.	Efficiency and High estimation for incomplete data sets	Slow convergence rate when data dimensionality is increased			✓	✓		✓	✓	✓	✓
[68]		Expectation–Maximization algorithm	Structured	Evaluate the selection and delivery of SaaS according to user needs using a SaaS quality model.	Automation of complete SaaS quality assessment.	Slow convergence and generally towards the local optima.	✓	✓	✓	✓			✓	✓	✓
[69]		Penalty-based Genetic Algorithm	Structured	Efficient placement of SaaS components considering the storage of their data in the Cloud	Optimization of performance and use of SaaS resources. Excellent scalability	Centralization. Parallel increase in computing time with the size of the network	✓		✓	✓		✓	✓	✓	✓
[70]		A Penalty-based Grouping Genetic Algorithm	Structured	Optimization of the dynamic management of resource allocation to composite SaaS in the Cloud.	High performance and scalability, reduced cost of used resources, minimal migration of VMs.	Long algorithm computation time.	✓	✓			✓	✓	✓		
[71]		Grouping Genetic Algorithm	Structured	Maintain composite SaaS performance through dynamic and optimal management of resource allocation in cloud data centers.	High efficiency and performance, minimizing the use of SaaS resources.	Long calculation time.	✓		✓	✓	✓			✓	✓

IV. CONCLUSION

Cloud Computing has become the leading IT platform for sharing hardware and software resources that are provided "as a service" on demand to end users via the Internet. SaaS is a type of Cloud service that has emerged as an effective reuse paradigm to offer hosted software solutions as a service to consumers who benefit from: eliminating software costs and software maintenance, Internet accessibility, high availability and per-use pricing. While the use of SaaS has increased throughout the IT world, improving and ensuring the quality of SaaS is becoming an important activity for successful SaaS management. Data Mining has become very important in the Cloud Computing industry. Through Data Mining, models and relationships and meaningful knowledge can be extracted from raw data which helps to make decisions and discover future trends and behaviors. This article provides an overview of the various Data Mining techniques used in Cloud Computing including: classification, regression, time series analysis, prediction, clustering, summarization, association rules and sequence discovery. The research also explains how Data Mining and Cloud Computing, particularly the SaaS

model, can be used reciprocally. Here, we have reviewed various approaches in the field, different classification and clustering algorithms have been used to support the Cloud environment. As a research observation, it should be noted that research that addresses the integration of Data Mining into the SaaS model is still limited and focuses on assessing the quality of SaaS services rather than improving the SaaS solutions provided.

For future work, we aim to propose an approach that uses Data Mining techniques in the SaaS model to maintain key characteristics of SaaS and improve the quality of services provided to users. In other words, the guarantee of a qualified service must be maintained in order to build a more complete and reliable Cloud environment.

ACKNOWLEDGMENT

This study was presented orally as abstract paper at the ICONDATA 2020 conference.

REFERENCES

[1] S. Bandela, R. Gadde and S. Pabboju, "Survey on Cloud Computing Technologies and Security Threats", International journal of engineering research and technology, Volume 2, Issue 6, May 2015.

[2] F. Shahzad, "State-of-the-art Survey on Cloud Computing Security Challenges, Approaches and Solutions", The 6th International Symposium on Applications of Ad hoc and Sensor Networks, PP. 357 – 362, 2014.

[3] https://www.nist.gov/

[4] R.A. Dhote, S. P. Deshpande, " Data Mining with Cloud Computing: - An Overview", International Journal of Advanced Research in Computer Engineering & Technology, Volume 5, Issue 1, January 2016.

[5] H. Ahmed, "Data Mining in Cloud Computing", International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015.

[6] A. Prasanth, "Cloud Computing Services: A Survey", International Journal of Computer Applications, Volume 46- No.3, May 2012.

[7] Q. Zhang, L. Cheng and R. Boutaba, "Cloud Computing: state-of-the-art and research challenges", Journal of Internet Services and Applications, Volume 1, Issue 1, pp 7–18, April 2010.

[8] X. Geng, Z. Yang, "Data Mining in Cloud Computing", Atlantis Press, 2013.

[9] J. ZENG, "The development and application of Data Mining based on Cloud Computing", First International Conference on Advanced Algorithms and Control Engineering, 2018.

[10] B. Kaur, "Software As A Service: A Brief Study", International Research Journal of Engineering and Technology, Volume: 02, Issue: 03, June-2015.

[11] H. I. Syed and N. A. Baig, "Survey On Cloud Computing", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4, April 2013.

[12] D. Jagli and A. Gupta, "Clustering Model for Evaluating SaaS on the Cloud", International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 12, December 2013.

[13] M.H. Parekh, "Enhancement Clustering of Cloud Datasets using Improved Agglomerative Technique", International Journal of Advanced Networking Applications, 2014.

[14] T.C. Sandanayake and P.G.C.Jayangani, "Current Trends in Software as a Service (SaaS)", International Journal for Innovation Education and Research, Volume: 6, No-02, pp.221-234, 2018.

[15] R.A. Kautkar, "A Comprehensive Survey on Data Mining", International Journal of Research in Engineering and Technology, Volume: 03, Issue: 08, August-2014.

[16] N. Jain and V. Srivastava, " Data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology, Volume: 02, Issue: 11, 2014.

[17] www.gartner.com

[18] B. Ambulkar and V. Borkar, "Data Mining in Cloud Computing", International Journal of Computer Applications, 2012.

[19] S. Mukherjee, R. Shaw, N. Haldar and S. Changdar, "A Survey of Data Mining Applications and Techniques", International Journal of Computer Science and Information Technologies, Vol. 6, PP.4663-4666, 2015.

[20] A.Sharma, R. Sharma,V.K. Sharma and V. Shrivatava, "Application of Data Mining - A Survey Paper", International Journal of Computer Science and Information Technologies, Vol. 5, PP.2023-2025, 2014.

[21] C. Mehta, "Basics of Data Mining: A Survey Paper", International Journal of Trend in Research and Development, Volume 4, 2017.

[22] A. Karahoca, D. Karahoca and M. Şanver, "Survey of Data Mining and Applications (Review from 1996 to Now)", 2012.

[23] https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/

[24] R. Kabilan and N. Jayaveeran, "Survay of Data Mining Techniques in Cloud Computing", International Journal of Scientific Engineering and Applied Science - Volume-1, Issue-8, November 2015.

[25] R.Ş. PETRE, "Data Mining in Cloud Computing", Database Systems Journal, volume 3, 2012.

[26] R. Ying, L. Hong, L. Hua-wei, Z. Li-jun and W. Li-na, " Data Mining Based on Cloud-Computing Technology", MATEC Web of Conferences, January 2016.

[27] C. Kaushal, A. Arya and S. Pathania, "Integration of Data Mining in Cloud Computing", Advances in Computer Science and Information Technology, Volume 2, Number 7, pp 48 – 52, 2015.

[28] D. Talia and P. Trunfio, "How Distributed Data Mining Tasks can Thrive as Knowledge Services", Communications of the ACM, vol. 53, n. 7, pp. 132-137, July 2010.

[29] U.S. Patki, "Clustering Algorithms in Cloud Computing Environment", International Research Journal of Computer Science, Volume 4, Issue 04, April 2017.

[30] A. Aparajita, S. Swagatika and D. Singh, "Comparative Analysis of Clustering Techniques in Cloud For Effective Load Balancing", International Journal of Engineering and Technology, pp. 47-51, 2018.

[31] P. Madhuri and I.K. Rajani, "Improve Performance of clustering on Cloud Datasets using improved Agglomerative CURE Hierarchical Algorithm", International Journal of Science, Engineering and Technology Research, Volume 4, Issue 6, June 2015.

[32] K. Srivastava, R. Shah, D. Valia, and H. Swaminarayan, "Data Mining Using Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment", International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013.

[33] M. Shindler , A. Wong , "Fast and Accurate k-Means For Large Datasets", 2011.

[34] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam , "Implementation of K-Means Clustering in Cloud Computing Environment", Research Journal of Applied Sciences, Engineering and Technology, PP.1391-1394, 2012.

[35] R. Atan, "Service Availability and Accessibility of Requirements Using Clustering in Cloud Environment", International Journal on New Computer Architectures and Their Applications, Volume 6, Issue 2, PP. 457-463, 2012.

[36] E. Sarkar, C.H. Sekhar , "Organizing Data in Cloud using Clustering Approach", International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May-2014.

[37] R. Asnani,"A distributed k-mean clustering algorithm for Cloud Data Mining", International Journal of Engineering Trends and Technology, Volume 30, 2015.

[38] B. Panchal and R.K.Kapoor, "Performance Enhancement of Cloud Computing using Clustering", International Journal of Computer Science and Network Security, Volume 14, June 2014.

[39] R.S. Sajjan and R.Y. Biradar, "Load Balancing using Cluster and Heuristic Algorithms in Cloud Domain", Indian Journal of Science and Technology, Vol 11, April 2018.

[40] Y.H.P. Raju and N. Devarakonda,"Cluster based Hybrid Approach to Task Scheduling in Cloud Environment", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 4, 2019.

[41] R. M. Esteves, R. Pais and C. Rong, "K-means Clustering in the Cloud - A Mahout Test", IEEE Workshops of International Conference on Advanced Information Networking and Applications, Singapore, 2011, pp. 514-519.

[42] S. Liu and Y. Cheng, "Research on K-Means Algorithm Based on Cloud Computing," *2012 International Conference on Computer Science and Service System*, Nanjing, 2012, pp. 1762-1765.

[43] Cui, X., Zhu, P., Yang, X. *et al*. Optimized big data K-means clustering using MapReduce. J Supercomput 70**,** 1249–1259 (2014).

[44] Yang, X., & Liu, P. A New Algorithm Of The Data Mining Model In Cloud Computing Based On Web Fuzzy Clustering Analysis, Journal of Theoretical and Applied Information Technology, Vol. 49 No.1, March 2013.

[45] M. Arjmand and F. Adibnia, "A Fuzzy KNN Classifier for Confidentiality of Cloud Tasks", International Journal of Humanities and Cultural Stadies, 2016.

[46] J. Wang, "A Novel K-NN Classification Algorithm for Privacy Preserving in Cloud Computing", Research Journal of Applied Sciences, Engineering and Technology, PP.4865-4870, 2012.

[47] P. Bajare, M. Bhoyate,Y. Bhujbal, E. Monika and V. Shinde, "k-Nearest Neighbor Classification Over Encrypted Cloud Data", Journal of Computer Engineering, PP. 45-48, 2015.

[48] V. Goutham, P. A. Reddy and K. Sunitha, " K-Nearest Neighbor Classification on Data Confidentiality and, Privacy of User's Input Queries", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 7, July 2016.

[49] R. Kour , S. Koul and M. kour , "A Classification Based Approach For Data Confidentiality in Cloud Environment", International Conference on Next Generation Computing and Information Systems, 2017.

[50] K. Patel and R. Srivastava, " Classification of Cloud Data using Bayesian Classification", International Journal of Science and Research, Volume 2, Issue 6, June 2013.

[51] F. Ebadifard and S.M. Babamir, "Dynamic task scheduling in Cloud Computing based on Naïve Bayesian classifier", International Conference on Information Technology, 2017.

[52] L. Zhou, H. Wang and W. Wang, "Parallel Implementation of Classification Algorithms Based on Cloud Computing Environment", TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol.10, pp. 1087-1092, September 2012.

[53] He Q., Zhuang F., Li J., Shi Z, "Parallel Implementation of Classification Algorithms Based on MapReduce". In: Yu J., Greco S., Lingras P., Wang G., Skowron A. Rough Set and Knowledge Technology. RSKT 2010. Lecture Notes in Computer Science, vol 6401. Springer, Berlin, Heidelberg

[54] A.B. Kamdar and J.M. Jagani, "A survey: classification of huge Cloud Datasets with efficient Map - Reduce policy", International Journal of Engineering Trends and Technology, Volume 18, 2014.

[55] F.O. Catak and M.E. Balaban , "CloudSVM : Training an SVM Classifier in Cloud Computing Systems", *ICPCA/SWS*, 2013.

[56] L. Shuhong, "Improved SVM in Cloud Computing Information Mining", International Journal of Grid Distribution Computing, Volume 8, pp.33-40, 2015.

[57] P. Kumar and A. Verma, "Scheduling Using Improved Genetic Algorithm in Cloud Computing for Independent Tasks", International Conference on Advances in Computing, Communications and Informatics, 2012.

[58] K. Zhu, H. Song, L. Liu, J. Gao and G. Cheng, " Hybrid Genetic Algorithm for Cloud Computing Applications", IEEE Asia-Pacific Services Computing Conference, Jeju, Korea (South), 2011.

[59] T.D. Le , V. Kantere and L. d' Orazio, " An efficient multi-objective genetic algorithm for Cloud Computing: NSGA-G", *IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, pp. 3883-3888, 2018.

[60] Z. Lijuan and Z. Shuguang, "The Strategy of Classification Mining Based on Cloud Computing", 1st International Workshop on Cloud Computing and Information Security, 2013.

[61] R. Latif, H. Abbas, S. Latif and A. Masood, " EVFDT: An Enhanced Very Fast Decision Tree Algorithm for Detecting Distributed Denial of Service Attack in Cloud-Assisted Wireless Body Area Network", Mobile Information Systems, 2015.

[62] V. Kanagalakshmi V and R. Gnanaselvam , "Review of Clustering Technique using SaaS on the Cloud", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 2, Issue 4, 2017.

[63] D. Jagli, S. Mahajan and N. S. Chandra, "CBC Approach for Evaluating Potential SaaS on the Cloud", International Technological Conference (I-TechCON), 2014.

[64] D. Jagli, S. Mahajan and N. S. Chandra, "Comparative Clustering Approach Intended for Evaluating SaaS", in International Journal of Computer Applications, Volume 169 – No.8, July 2017.

[65] D. Jagli, S. Mahajan and N. S. Chandra, "Implementation of Pam Cluster for Evaluating SaaS on the Cloud Computing Environment", Journal of Engineering, Vol. 08, Issue 4, PP 84-88, April 2018.

[66] R. Kamalraj, A.R. Kannan, S.Vaishnavi and V. Suganya, "A DataMining BasedApproach for Introducing Products in SaaS (Software as a Service)", International Journal of Engineering Innovation & Research, Volume 1, Issue 2, 2012.

[67] D.Jagli, S. Purohit and N.S. Chandra, "EM Clustering Model for Evaluating SaaS on Cloud Computing Environment", Journal of Computer Engineering, Volume 20, Issue 2, PP 67-72, 2018.

[68] D.Jagli, S. Purohit and N.S. Chandra, "SAASQUAL: A Quality Model for Evaluating SaaS on the Cloud Computing Environment", In Big Data Analytics, pp. 429-437. Springer, Singapore, 2018.

[69] Z. I. M. Yusoh and M. Tang, "A Penalty-based Genetic Algorithm for the Composite SaaS Placement Problem in the Cloud", IEEE World Congress on Computational Intelligence, Barcelona, pp. 1-8, July 2010.

[70] Z. I. M. Yusoh and M. Tang, "Clustering composite SaaS components in Cloud Computing using a Grouping Genetic Algorithm," *IEEE Congress on Evolutionary Computation*, Brisbane, QLD, pp. 1-8, 2012.

[71] Z. I. M. Yusoh and M. Tang, "A penalty-based grouping genetic algorithm for multiple composite SaaS components clustering in Cloud," *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Seoul, pp. 1396-1401, 2012.