International Journal of Cloud-Computing and Super-Computing
Vol. 3, No. 1, (2016), pp.13-24
http://dx.doi.org/10.21742/ijcs.2016.3.1.03

QoS based Machine Learning Algorithms for Clustering of Cloud Workloads: A Review

Sukhpal Singh[1*] and Inderveer Chana[2]

[1,2]Computer Science and Engineering Department, Thapar University, Patiala, Punjab, India-147004
[1]ssgill@thapar.edu, [2]inderveer@thapar.edu

Abstract

Data mining, known as knowledge discovery is a computer-assisted, analytical process of digging through and analyzing large amount of data and extracting knowledge from the data. Data mining technologies has made many industries like marketing, sales, healthcare organization, financial institutions etc. quite successful. It has a lot of benefits in various fields. It helps in the quick analysis of data and has improved the quality of decision making process and is used to turn information into actionable knowledge. In this paper, data mining process and knowledge discovery process is discussed. Three well known data mining classifier algorithms namely ID3, J48 and Naive Bayes are discussed and their performance has been evaluated using different parameters to find the best algorithm. Further, Naive Bayes classifier algorithm is used for classification of workloads based on different Quality of Service (QoS) parameters.

Keywords: *Cloud Computing, Cloud Workloads, Machine Learning, Data Mining, Knowledge Discovery Process, Quality of Service*

1. Introduction

Extraction of trends and patterns from data manually has occurred for centuries. Bayes Theorem and regression analysis were used earlier for carrying out manual extraction. The increase in power of computer technology has brought a momentum in data collection, storage and ability of manipulating the data [1]. With the growth in size of datasets and its complexity, the manual processing of data has been augmented with automated data processing using neural networks, decision trees, cluster analyses and support vector machines. Thus data mining can be defined as process of applying these techniques to discover hidden trends and patterns in the data [2]. Data mining is comparatively a unique process. As in standard operations on database, the results provided to the user are something that the user already knew to be existed. But data mining on the other hand extracts and provides information that the user did not know like the relationship between customer behavior and the variables that are non-intuitive. And as the information is not known beforehand it is a bigger leap to take out the result of the system and use it as a solution to some business problem [3]. Data mining tools help in predicting future trends and behaviors and allow business to make knowledge-driven and proactive decisions. Business questions that were earlier time consuming to resolve can be answered efficiently and in much less time using data mining tools. These tools analyze datasets to find hidden patterns and predictive information that can be missed by experts as it lies beyond their expectations.

Print ISSN: 2205-8567, eISSN: 2207-399X IJCS

Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data [4]. The knowledge discovery process is described in Figure 1.

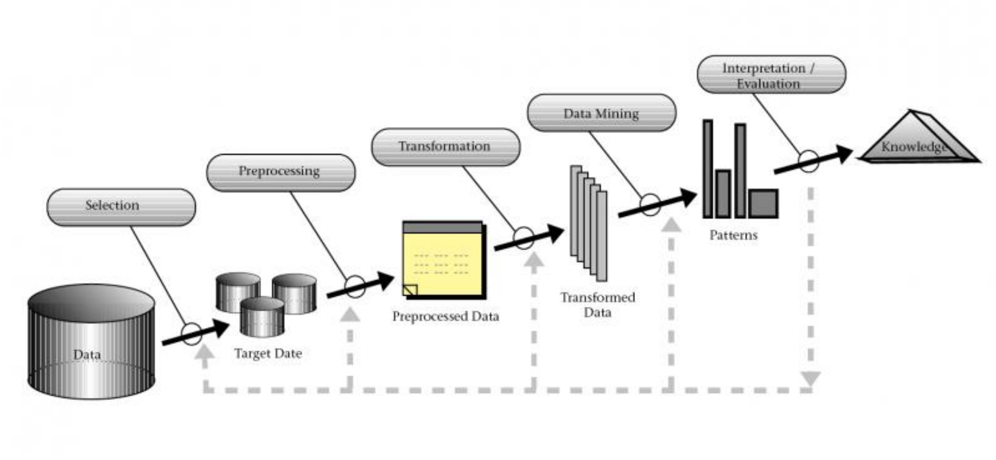


Figure 1. Knowledge Discovery Process

Through computerization, data mining process provides way to make better use of data. Data mining software use modeling techniques using mathematical relations based on data where answer is known and then the same model is applied to other data where answer is not known or hidden [5]. Phases in process of data mining are shown in Figure 2. Data comes from variety of sources is integrated into a single data store called target data. Data then is pre-processed and transformed into standard format. The data mining algorithms process the data to the output in form of patterns or rules. Then those patterns and rules are interpreted to new or useful knowledge or information.

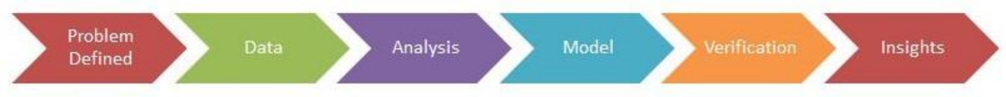


Figure 2. Knowledge Discovery Process

Data mining techniques are being used for making decisions based on some specified rules. The three well known data mining classifier algorithms namely ID3, J48 and Naive Bayes are used for data mining [6] [7]. Among them Naive Bayes is one of the most effective inductive learning algorithms used to make decisions based on predefined rules in decision trees is concerned.

1.1. Our Contributions

We have presented data mining process and knowledge discovery process in this research paper. This research work is an extension of our previous research work [9] [10]. Three well known data mining classifier algorithms namely ID3, J48 and Naive Bayes is discussed and their performance has been evaluated using different parameters to find the best algorithm. Further, Naive Bayes classifier algorithm is used for classification of workloads based on different Quality of Service (QoS) parameters.

The motivation of this research work is to evaluate the performance of different data mining algorithms to identify the best algorithm for clustering and further that algorithm is used for clustering of workloads based on different QoS parameters. The organization of rest of this paper is as follows: Sec. 2 presents state of the art of machine learning algorithms. Sec. 3 describes the Naive Bayes classifier algorithm based workloads clustering. Sec. 4 presents the conclusion and future scope of this research work.

2. Machine Learning Algorithms: State-of-the-Art

Systems that create classifiers are one of the frequently used tools in data mining. Such systems take a group of different cases as input; every one belongs to a small amount of classes defined by a stable set of attributes and output a classifier that can truthfully forecast the class to which a new case belongs. Datasets can have nominal, numeric or mixed attributes and classes. Not all classification algorithms perform well for different kinds of attributes, classes and for datasets of different dimensions. In order to design a standard classification tool, one should consider the behavior of various existing classification algorithms on different datasets. The choice of a classification algorithm depends on the desires and the nature of classification required. Literature reported [1-10] [22-26] the following machine learning algorithms:

2.1. Decision Trees

Decision tree has been considered in details in both areas of pattern recognition and machine learning. This creates the experience expanded by individuals working in the region of machine learning and describes a computer program called ID3, which has evolved to a new system, named C4.5 (an enhanced version of C4.5 is J48).

2.2. ID3 Algorithm

ID3 is one of the well-known Inductive Logic Programming procedures. It is basically an attribute based machine-learning algorithm that builds a decision tree based on a training set of data and an entropy measure to build the leaves of the tree. The informal formulation of ID3 is as follows:

- Define the element that has the highest statistics gain on the training set.
- Use this element as the root of the tree; generate a branch for each of the values that the attribute can take.
- For every branch, repeat this process with the subset of the training set that is categorized by this branch.

2.3. J48 Algorithm

J48 (enhanced version of C4.5) is developed based on the ID3 algorithm, with additional features to address problems that ID3 was unable to deal. In exercise, C4.5 uses one successful process for finding high correctness guesses, based on snipping the instructions dispensed from the tree constructed during the learning phase. Conversely, the principal disadvantage of C4.5 rule sets is the amount of CPU time and memory they need. Given a set S of cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the different cases in C belong to the similar class or C is small, the tree is leaf labelled with the most frequent class in C.

- Otherwise, select a test based on a single attribute with two or more results. Create this test as the root of the tree with one branch for every result of the test, partition C into corresponding subclasses C1, C2… according to the result for every different case, and apply the same technique recursively to every subclass.

There are usually many tests that could be chosen in this last step. J48 uses two heuristic criteria to rank possible tests: statistics gain, which reduces the total entropy of the subclasses {Ci} and the default gain ratio that distributes statistics gain by the information provided by the test outcomes. After the building process, each attribute test along the path from the root to the leaf becomes a rule antecedent (precondition) and the classification at the leaf node becomes the rule consequence (post condition). To illustrate the post pruning of the rules, let us consider the following rule generated from the tree:

***IF** Condition **THEN** Conclusion*

This rule is pruned by removing any antecedent whose removal does not worsen its estimated accuracy.

2.4. Naive Bayes Algorithm

Naive Bayes is an extension of Bayes theorem in that it assumes independence of attributes. This supposition is not stringently accurate when considering grouping based on text extraction from a document as there are relationships between the words that collect into concepts. Problems of this kind, called problems of supervised classification, are ubiquitous. It is simple to construct without any requirement for complex iterative parameter approximation patterns. This means it may be enthusiastically applied to huge data sets. It is robust, easy to interpret, and often does surprisingly well though it may not be the best classifier in any particular application.

2.4.1. Experiment Results: The implementation of proposed algorithm has been done through the MATLAB [27]. The Figure 3 shows the experimental results using decision trees with the ID3 and J48 (extension of C4.5) algorithms, along with the results obtained from Naive Bayes algorithm.

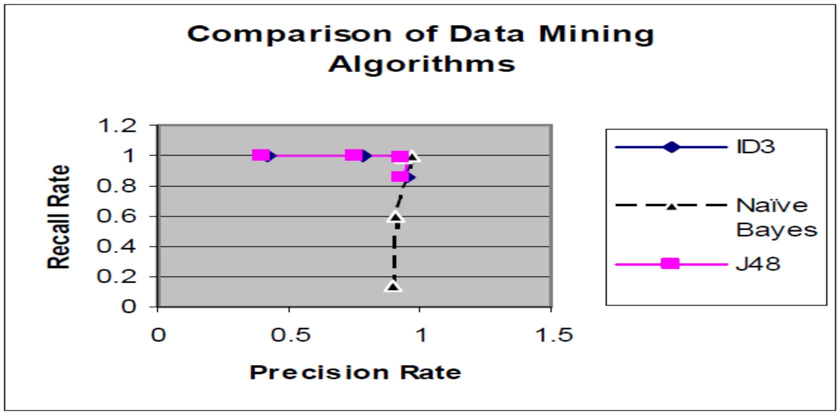


Figure 3. Precision-Recall Characteristics

Precision Rate - It is the fraction of retrieved Information that is relevant to the search.

$$\text{Precision Rate} = \frac{\text{Relevant Information U Retrieved Information}}{\text{Retrieved Information}}$$

Recall Rate - Recall in information retrieval is the fraction of the Information that are relevant to the query that are successfully retrieved.

$$\text{Recall Rate} = \frac{\text{Relevant Information U Retrieved Information}}{\text{Relevant Information}}$$

Error rate - The degree of errors encountered during extract information from a data set and transforms it into an understandable structure.

In Table 1, comparison is done with respect to time taken to build the model, along with the error rate for all the three data mining algorithms.

Table 1. Comparison of Data Mining Algorithms

Experiments	Overall Error Rate	Time taken to build the Model in Seconds
Naive Bayes	3.46%	0.11
J48	3.47%	0.88
ID3	3.47%	1.8

Comparison of data mining algorithms, as shown in Figure 4 suggest that the ability of the various classification methods examined could be considered as good, i.e. the classifier stability is very strong. Root mean square error in the same way suggests that the error rate is very small, which can be considered as a measure of effectiveness of the model. It can be observed from the Figure 4, that the accuracy of overall classification for Naive Bayes for all classes is better than the overall accuracy obtained in the case of J48 and ID3 algorithms. It is observed from the results obtained by experimentation that the Naive Bayes model is quite appealing because of its simplicity, elegance, robustness and effectiveness.

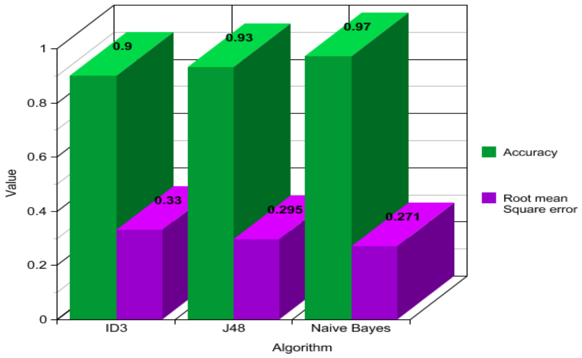


Figure 4. Comparison of Data Mining Algorithms

In our earlier work [9] [10] [13] [15] [17], we have identified various research issues related to QoS and SLA for cloud resource scheduling and have developed a QoS based resource provisioning technique (Q-aware) to map the resources to the workloads based on used requirements described in the form of SLA. Further, resource scheduling framework (QRSF) has been proposed, in which provisioned resources have been scheduled by using different resource scheduling policies (cost, time, cost-time and bargaining based). The concept of QRSF has been further extended by proposing energy-aware autonomic resource scheduling technique (EARTH), in which IBM's autonomic computing concept has been used to schedule the resources automatically by optimizing energy consumption and resource utilization where user can easily interact with the system using available user interface. In this work, performance of different data mining algorithms is evaluated to identify the best algorithm for clustering and further that algorithm is used for clustering of workloads based on different QoS parameters.

3. Naive Bayes Classifier Algorithm based Workloads Clustering

Workload clustering is done based on Naive Bayes classifier algorithm. Table 2 shows the different type of workloads and their QoS requirements considered in this research work [11-21].

Table 2. Different type of workloads and their QoS requirements [9] [10]

Id	Workload	QoS Requirements
W1	**Web sites**	➢ Reliable storage ➢ High network bandwidth ➢ High availability
W2	**Technological Computing**	➢ Computing capacity
W3	**Endeavour Software**	➢ Security ➢ High availability ➢ Customer confidence Level ➢ Correctness
W4	**Performance Testing**	➢ Computing capacity
W5	**Online Transaction Processing**	➢ Security ➢ High availability, ➢ Internet accessibility ➢ Usability
W6	**E-Com**	➢ Variable computing load ➢ Customizability
W7	**Central Financial Services**	➢ Security ➢ High availability ➢ Changeability ➢ Integrity
W8	**Storage and Backup Services**	➢ Reliability ➢ Persistence
W9	**Productivity Applications**	➢ Network bandwidth ➢ Latency ➢ Data backup ➢ Security
W10	**Software/Project Development and Testing**	➢ User self-service rate ➢ Flexibility ➢ Creative group of infrastructure services

		➢ Testing time
W11	**Graphics Oriented**	➢ Network bandwidth ➢ Latency ➢ Data backup ➢ Visibility
W12	**Critical Internet Applications**	➢ High availability ➢ Serviceability ➢ Usability
W13	**Mobile Computing Services**	➢ High availability ➢ Reliability ➢ Portability

The relation between quality attributes and cloud workloads has been identified using Naive Bayes classifier algorithm and described in the form of Matrix as shown in Table 3.

Table 3. Quality Attributes and Cloud Workloads Matrix

Quality Attributes	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**	**W13**
Reliable Storage	√												
Latency									√		√		
Computing Capacity		√		√									
Data Backup									√		√		
Customer Confidence Level			√										
User Self-Service Rate										√			
Correctness			√										
Flexibility										√			
Security			√		√		√		√				
Internet Accessibility					√								
Creative Group of Infrastructure Services										√			
Usability					√							√	
Variable Computing Load						√							
Customizability						√							
Testing Time										√			
Changeability							√						
Visibility											√		
Integrity							√						
Reliability								√	√				√
Serviceability												√	
Persistence								√	√				
Portability													√
Performance				√									
High Network Bandwidth	√									√	√		
High Availability	√		√		√		√					√	√

Clustering of different cloud workloads is shown in Table 4.

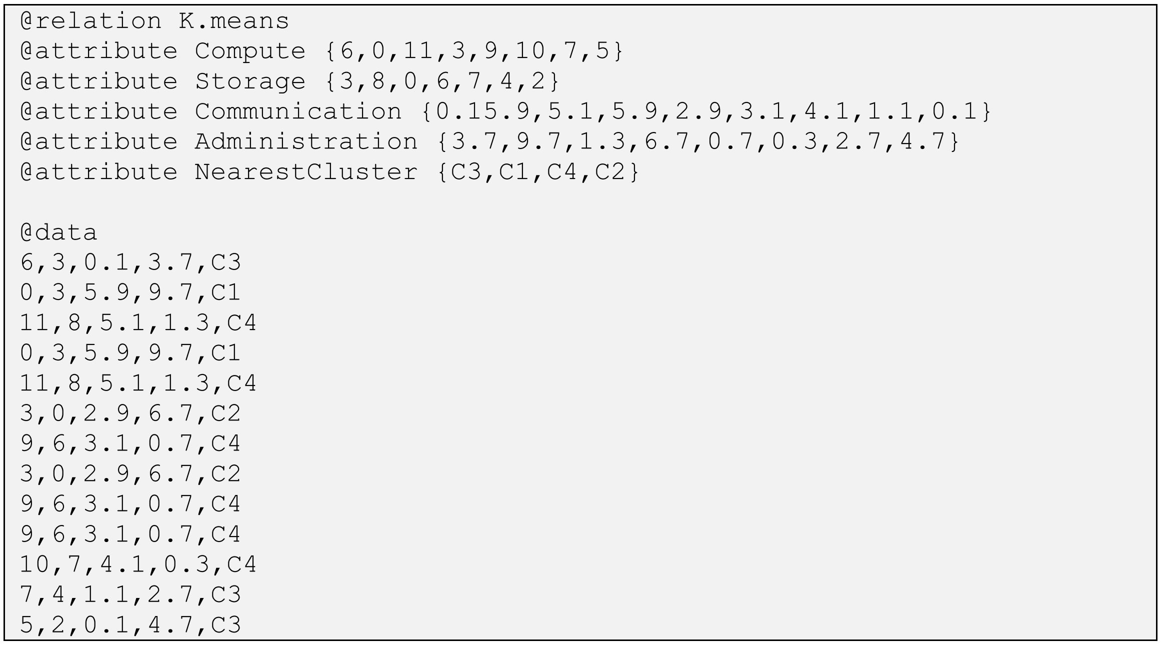
Table 4. Clustering of different cloud workloads [9] [10]

Cluster	Cluster Name	Workloads
C1	Compute	W2, W4
C2	Storage	W6, W8
C3	Communication	W1, W12, W13
C4	Administration	W3, W5, W7, W9, W10, W11

3.1. Experiment Results

The results of workloads clustering were taken by a data mining tool i.e. WEKA [7]. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of workloads sharing a set of clusters. An Attribute-Relation File Format file (k.means.arff) for cloud workloads clusters and their corresponding values are shown in Figure 5.

ARFF files have two distinct sections. The first section is the cluster information, which is followed the workload distance information. The Header of the ARFF file contains the name of the relation, a list of the clusters (named Compute, Storage, Communication and Administration), and their minimum distance value. These results are measured with the help of data mining tool i.e. WEKA Tool [7]. The results are taken through the use of WEKA tool. There are four different clusters (C1 = Compute, C2 = Storage, C3 = Communication and C4 = Administration) along with nearest cluster. In this algorithm, the minimum distance is calculated and put a particular workload into suitable cluster.



```
@relation K.means
@attribute Compute {6,0,11,3,9,10,7,5}
@attribute Storage {3,8,0,6,7,4,2}
@attribute Communication {0.15.9,5.1,5.9,2.9,3.1,4.1,1.1,0.1}
@attribute Administration {3.7,9.7,1.3,6.7,0.7,0.3,2.7,4.7}
@attribute NearestCluster {C3,C1,C4,C2}

@data
6,3,0.1,3.7,C3
0,3,5.9,9.7,C1
11,8,5.1,1.3,C4
0,3,5.9,9.7,C1
11,8,5.1,1.3,C4
3,0,2.9,6.7,C2
9,6,3.1,0.7,C4
3,0,2.9,6.7,C2
9,6,3.1,0.7,C4
9,6,3.1,0.7,C4
10,7,4.1,0.3,C4
7,4,1.1,2.7,C3
5,2,0.1,4.7,C3
```

Figure 5. Attribute-Relation File Format File

4. Conclusions and Future Directions

In this paper, different data mining techniques have been compared based on different criteria and the best data mining technique that has been used to make the decision to choose scheduling criteria has been discussed. Three well known data mining classifier algorithms namely ID3, J48 and Naive Bayes has been discussed and performance has been evaluated using MATLAB based on different parameters like precision rate, recall rate, error rate and time. Performance evaluation is shown that Naive Bayes classifier algorithm performs better. Naive Bayes classifier algorithm is used for classification of algorithms. WEKA tool is used to further elaborate the clustering of cloud workload based on different QoS parameters. Further based on clustering of workloads, resources can be scheduled in an effective way in cloud computing.

Acknowledgments

One of the authors, Sukhpal Singh, acknowledges the Department of Science and Technology (DST), Government of India, for awarding him the INSPIRE (Innovation in Science Pursuit for Inspired Research) Fellowship (Registration/IVR Number: 201400000761 [DST/INSPIRE/03/2014/000359]) to carry out this research work. We would like to thank all the anonymous reviewers for their valuable comments and suggestions for improving the paper. We would also like to thank Dr. Maninder Singh [EC-Council's Certified Ethical Hacker (C-EH)] for his valuable suggestions.

References

[1] M. J. A. Berry and G. Linoff, "Mastering data mining", The Art and Science of Customer Relationship Management, **(1999)**.

[2] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, **(2003)**.

[3] J. R. Quinlan, "C4.5: Programs for machine learning", Morgan Kaufmann, **(1993)**.

[4] M. Umano, H. Okamoto, I. Hatono, H. Tamura, F. Kawachi, S. Umedzu, and J. Kinoshita, "Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems" Proceedings of the Third IEEE Conference on World Congress on Computational Intelligence and Fuzzy Systems, **(1994)**, pp. 2113-2118.

[5] J. R. Quinlan, "Decision trees and decision making", IEEE transaction on system Man cyber, vol. 20, no. 2, **(1990)**, pp. 339-346.

[6] M. Panda and M. R. Patra, "A comparative study of data mining algorithms for network intrusion detection", Proceedings of the Emerging Trends in Engineering and Technology, ICETET, **(2008)**.

[7] M. Hall, E. Frank, G. Holmes , B. Pfahringer and I. . H. Witten, "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, **(2009)**, pp. 10-18.

[8] S. Singh and I. Chana, "Cloud Based Development Issues: A Methodical Analysis", International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 2, no. 1, **(2012)**, pp. 73-84.

[9] S. Singh and I. Chana, "Q-aware: Quality of Service based Cloud Resource Provisioning", Computers & Electrical Engineering - Journal – Elsevier, **(2015)**. DOI: http://dx.doi.org/10.1016/j.compeleceng.2015.02.003

[10] S. Singh and I. Chana, “QRSF: QoS-aware resource scheduling framework in cloud computing”, The Journal of Supercomputing, vol. 71, no. 1, **(2015)**, pp. 241-292.

[11] S. Singh and I. Chana, “Consistency verification and quality assurance (CVQA) traceability framework for SaaS”, Proceeding of the IEEE 3rd International on Advance Computing Conference (IACC), **(2013)**, pp. 1-6.

[12] S. S. Gill, “Autonomic Cloud Computing: Research Perspective”, **(2015)**, pp. 1-3. Retrieved from http://arxiv.org/ftp/arxiv/papers/1507/1507.01546.pdf

[13] S. Singh and I. Chana, “EARTH: Energy-aware Autonomic Resource Scheduling in Cloud Computing” Journal of Intelligent and Fuzzy Systems, **(2015)**, pp. 1-16.

[14] S. Singh and I. Chana, “Introducing Agility in Cloud Based Software Development through ASD”, International Journal of u-and e-Service, Science and Technology, vol. 6, no. 5, **(2013)**, pp. 191-202.

[15] S. Singh and I. Chana, “Advance billing and metering architecture for infrastructure as a service”, International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 2, no. 2, **(2013)**, pp. 123-133.

[16] S. Singh and I. Chana, “QoS-aware Autonomic Cloud Computing for ICT”, Proceeding of the International Conference on Information and Communication Technology for Sustainable Development (ICT4SD - 2015), Springer International Publishing, **(2015)**.

[17] S. Singh and I. Chana, “QoS-aware Autonomic Resource Management in Cloud Computing: A Systematic Review”, ACM Computing Surveys, vol. 48, no. 3, **(2015)**, pp. 1-39.

[18] S. Singh and I. Chana, “Energy based efficient resource scheduling: a step towards green computing”, International Journal of Energy Information and Communications, vol 5, no. 2, **(2014)**, pp. 35-52.

[19] S. Singh and I. Chana, “Formal Specification Language Based IaaS Cloud Workload Regression Analysis”, arXiv preprint arXiv:1402.3034, **(2014)**. Retrieved from http://arxiv.org/ftp/arxiv/papers/1402/1402.3034.pdf

[20] S. Singh and I. Chana, “Cloud Resource Provisioning: Survey, Status and Future Research Directions”, Knowledge and Information Systems, **(2015)**.

[21] S. Singh and I. Chana, “A Survey on Resource Scheduling in Cloud Computing Issues and Challenges”, Journal of Grid Computing, **(2015)**.

[22] D. Sarddar, E. Nandi, R. K. Gupta, R. K. Pateriya, A. Abdulmohson, S. Pelluri, and R. Sirandas, “Implement of Dynamic Time Quantum Shortest Load First Scheduling for Efficient Load Balancing” International Journal of Cloud-Computing and Super-Computing, vol. 2, no. 1, **(2015)**, pp. 1-8.

[23] R. K. Gupta and R. K. Pateriya, “Energy Efficient Virtual Machine Placement Approach for Balanced Resource Utilization in Cloud Environment”, International Journal of Cloud-Computing and Super-Computing, vol. 2, no. 1, **(2015)**, pp. 9-20.

[24] A. Abdulmohson, S. Pelluri, and R. Sirandas, “Energy Efficient Load Balancing of Virtual Machines in Cloud Environments”, International Journal of Cloud-Computing and Super-Computing, vol. 2, no. 1, **(2015)**, pp. 21-34.

[25] V. S. Rathor, R. K. Pateriya, R. K. Gupta, M. Shelar, S. Sane, V. Kharat, and R. Jadhav, “An Efficient Virtual Machine Scheduling Technique in Cloud Computing Environment”, International Journal of Cloud-Computing and Super-Computing, vol. 1, no. 1, **(2015)**, pp. 1-14.

[26] M. Shelar, S. Sane, V. Kharat, and R. Jadhav, “Efficient Virtual Machine Placement with Energy Saving in Cloud Data Center”, International Journal of Cloud-Computing and Super-Computing, vol. 1, no. 1, **(2014),** pp. 15-26.

[27] M. Grant, S. Boyd and Y. Ye, "CVX: Matlab software for disciplined convex programming", **(2008)**.

Author's Biography

Sukhpal Singh obtained the Degree of Master of Engineering in Software Engineering from Thapar University, Patiala. Mr. Singh received the Gold Medal in Master of Engineering in Software Engineering. Presently he is pursuing Doctoral degree in Cloud Computing from Thapar University, Patiala. Mr. Singh is on the Roll-of-honor being DST Inspire Fellow as a SRF Professional. He has done certifications in Cloud Computing Fundamentals, including Introduction to Cloud Computing and Aneka Platform (US Patented) by ManjraSoft Pty Ltd, Australia and Certification of Rational Software Architect (RSA) by IBM India. His research interests include Software Engineering, Cloud Computing, Operating System and Databases. He has more than 25 research publications in reputed journals and conferences.

Inderveer Chana joined Computer Science and Engineering Department of Thapar University, Patiala, India, in 1997 as Lecturer and is presently serving as Professor in the department. She is Ph.D. in Computer Science with specialization in Grid Computing and M.E. in Software Engineering from Thapar University and B.E. in Computer Science and Engineering. Her research interests include Grid and Cloud computing and other areas of interest are Software Engineering and Software Project Management. She has more than 100 research publications in reputed Journals and Conferences. Under her supervision, more than 30 ME thesis and four Ph.D thesis have been awarded and four Ph.D. thesis are on-going. She is also working on various research projects funded by Government of India.