Middle-East Journal of Scientific Research 21 (2): 376-384, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.21.02.21219

An Efficient Integrated Approach for Information Retrieval Using Fuzzy Artificial Bee Colony Optimization Based On Cloud Computing

K. Sathesh Kumar and M. Hemalatha

Department of Computer Science,
Karpagam University, Coimbatore, India

Abstract: Elicitation of hidden knowledgeable information from voluminous data is the most prompting technique in data mining (DM). This is proved by DM algorithms for knowledge discovery. Data mining incorporated with Cloud computing technology helps to achieve maximize profit and minimum cost with different possible ways through shared Cloud resource. This paper examines the development framework for information retrieval using DM techniques in Cloud computing environment. Here, the proposed work involves the four different approaches like Boosted K-NN, SVM, Fuzzy based ant colony optimization and Fuzzy artificial bee colony optimization for handling the cloud computing dataset. The main aim of this work is to deploy these approaches in Google Cloud using Google App Engine with Cloud SQL. The performance results show better outcomes compared to real world applications and also reduce the computing time, cost and infrastructure, to produce the results with mean time.

Key words: Cloud computing · Cloud SQL · Data mining · Artificial bee colony · Rule optimization

INTRODUCTION

Today's business world is fast and dynamic in nature where it involves a lot of data gathered from different perspectives and are stored in a data warehouse. The most demanding task of the business people is to transform these data into a useful format called knowledge/information. Data mining techniques are utilized to accomplish this task Cloud computing can be recognized as a new approach in Data Mining. There is a lot of information and tragically this immense measure of information is challenging to mine and examine with computational resources. With cloud computing, ideal model of data mining could be more receptive simply because of cost and computational resources. Here, we discuss the utilization of cloud computing platforms as a conceivable solution for mining and dissecting a large amount of information. The execution of data mining in Cloud Computing technology permits associations to centralize the administration of the Software and storage space, with the assurance of productive, dependable and secure administrations for their users [1-5].

In the context of cloud with data mining technique aspects, different researches have been made for performance enhancement. Deploying the K-means algorithm [6-8] in a cloud environment performs the retrieving information on extensive databases and archiving them with less expense. It focuses on revealing the structure of document collections, summarizing their content. Outcome of the experimental results while deploying this algorithm is it improves the computing time, scalability and performance. Retrieving information through agent framework, [9, 10] gives the guarantee for information with less expensive time and cost. Reducing the map technology framework in a cloud environment mainly focuses on reducing the search time and to get an efficient retrieval. In order to improve the execution time and computation cost for retrieving and extracting the required data from a cloud environment efficiently by parallel computing and decision tree algorithms. This work presents on the data mining algorithm implemented in on cloud environment, to improve the performance [11-13].

Artificial Bee Colony (ABC) algorithm is a kind of swarm intelligence used for solving combinational optimization problems. This algorithm is based on a

Corresponding Author: K. Sathesh Kumar, Department of Computer Science, Karpagam University, Coimbatore, India.

particular intelligence behavior of the swarm. This algorithm recently introduced optimization algorithm based on foraging behavior of the bee algorithm for solving rule optimization problems. The ABC algorithm mainly contains three groups; first one, employee bees; second is onlookers and scouts. First half of the colony consists of the employee bees and second half consists of onlooker bees. For every food source there is only one employee bees. In other words, the number of employee bees is equal to food source, the employee bee of an abounded food source become scouts. This work presents data mining algorithm implemented on cloud environment, optimizing the rules through Fuzzy Artificial Bee Colony optimization (FABCO) for reducing the primitive rules and generates the final set of rules. The above proposed method was compared with the traditional classification algorithms such as K-Nearest Neighbor (KNN), Particle Swarm Optimization (PSO) and Artificial Bee Colony Optimization (ABCO) but this proposed method FABCO out performs best classification accuracy among the rest. The above method was successfully deployed into Google cloud services using cloud SQL based on Platform as a Service (PaaS), while integrating the above data mining techniques in a cloud environment to reduce the cost, time and infrastructure

Background

Datamining: The procedure of concentrating functional data from different sources is reputed to be data mining. In people's view, greater part of the data mining is an equivalent word of Knowledge discovery. In any case data mining might be acknowledged as a stage of knowledge discovery in databases (KDD). KDD process incorporates data cleaning (to evacuate noise and conflicting information), data integration data selection (where different information sources may be consolidated), data selection (where information pertinent to the dissection undertaking are recovered from the database), data transformation (where information are converted or combined into structures for mining by performing outline or aggregation operations), data mining (a key process where canny systems are connected with a specific end goal to concentrate data designs), pattern evaluation (to distinguish the genuinely fascinating examples speaking to learning dependent upon a few interesting measures) and knowledge presentation (where visualization and information representation procedures are utilized to present the mined information to the user [14].

Cloud Computing: Cloud computing is an innovation that uses the web and central remote servers to maintain information and provisions. Cloud computing permits individual users and organizations to utilize requisitions without any establishment and access their particular indexes on any machine with web access. This technology takes into account significantly more effective registering by bringing together space, memory and preparing data transfer capacity [15].

Need for Cloud Storage: While using the database on the local storage the problem of accessing anywhere at anytime becomes a crucial problem in business activities. This problem can be competed by using a Cloud Database instance is beneficial for customers who want or need high performance and redundant storage, the ability to scale their instance and grow their storage as needed and the freedom to focus on the database itself and not the underlying infrastructure. A cloud database can be a traditional database such as a MySQL or SQL Server database that has been adopted for cloud use, A native cloud database such as Xeround's MySQL Cloud database tends to be better equipped to optimally use cloud resources and to guarantee scalability as well as availability and stability. Cloud databases can offer significant advantages over their traditional counterparts, including increased accessibility, automatic failover and fast automated recovery from failures, automated on-the-go scaling, minimal investment and maintenance of in-house hardware and potentially better performance. At the same time, cloud databases have their share of potential drawbacks, including security and privacy issues as well as the potential loss of or inability to access critical data in the event of a disaster or bankruptcy of the cloud database service provider.

Proposed Work: Figure 1 shows the four different stages of our proposed work in cloud computing based database. The dataset collected in real time applications are not always complete due to the presence of missing value which is high. To handle such problems, the boosted k-NN is utilized to impute the missing value efficiently. The second stage of handling voluminous dataset in data mining is another high complexion to increase the speed of technique; the potential attributes are identified using Support Vector Machine (SVM). The third stage is identification of patterns among the dataset using the clustering approach known as Fuzzy ant colony based optimization, the patterns recognized are generated

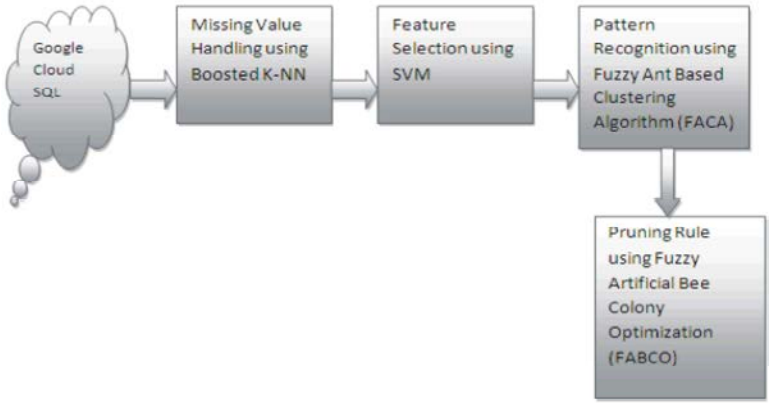


Fig. 1: Structure of the proposed framework

Fig. 2: Architecture of Google API service with Cloud SQL

as rules with the help of Fuzzy association rule mining for information retrieval. The rules generated using Fuzzy association rule mining are fine tuned with the help of Fuzzy artificial bee colony optimization FABCO to increase the performance of classification. Rules generated by Fuzzy association do not produce better results, thus FABCO which overcomes the challenge has been utilized.

Architecture of Google ApI Service With Cloud SQL: Figure 2 shows the Architecture of Google API service with Cloud SQL as described below.

Google App Engine (GAE): Google App Engine is different from most of the cloud system because of easy access, flexibility and low cost, easy deployment. Furthermore, it provides a development platform to create application under hosting platform. It is a pure PaaS cloud targeted traditional web application thereby enforcing a separation between a stateless computation tier and a tasteful storage tier. The Virtualization and the elasticity seem to be so visible in IaaS model but completely invisible here. Selling propositions of this model is automatic elasticity in terms of capacity requirement changes [16].

Google Cloud SQL: Google Cloud SQL is a fully-managed web service that enables creation, configuration, manages relational databases that relies on Google's infrastructure and maintains databases, allows focusing on the applications and services. It also offers MySQL database, Google Cloud SQL to move data and all kinds of applications within the cloud. This provides high data portability and helps to achieve a faster time to market, quickly leverage the existing database, while creating a Google Cloud SQL instance, it enables the user with synchronous or asynchronous replication for data. Google Cloud SQL is a kind of MySQL database that relies on Google's infrastructure. It has all possible capabilities and functionality of MySQL with a few additional features. It is easy to use and does not need any sort of software installation, maintenance and it is ideal for small- to medium-sized applications.

GAE Data Store: Archiving information in an adaptable web requisition could be unreliable. A client could be connecting with any of the many web servers at a given time and the client's next request could head off to an alternate web server than the past appeal. All web servers need to be connected with information that additionally spread out crosswise over many machines, potentially in diverse areas as far as possible.

Google Social Graph API: Google's Social Graph API is an administration taking advantage of "companion" associations between individuals' sites and web administrations. It is dependent upon XFN and FOAF, which essentially implies exceptional "companion" or "me" tags installed in standard ole' hyperlinks.

G- Data: G-data provides a basic convention for perusing and composing information on the Internet, scenario by Google. G-data joins together regular XML-based syndication designs with a feed distributed framework depending upon the Atom Publishing Protocol, in addition to a few extensions for handling queries.

MATERIALS AND METHODS

The proposed algorithms were implemented in Matlab to facilitate design and development of the application. To deploy the application in Google, the Google App Engine Plug-In is downloaded from Google's official site. To create Database and, table Google Cloud SQL with the help of Matlab JDBC toolbox is used.

Webservice in MATLAB: Figure 3 shows the Web service in Matlab Web administrations permit provisions running on divergent PCs, working frameworks and advancement situations to correspond with one another. Utilizing Web administrations technologies, client workstations can access and execute APIs residing on a remote server.

The client and server communicate via XML-formatted messages, emulating the W3cSoap convention and normally by means of the http protocol. Matlab acts as a Web administration customer, giving capacities access to the existing Web services on a server. The capacities expedite correspondence with the server, alleviating the need to work with Xml, complex Soap messages and exceptional Web administrations devices. Through these capacities, Web services can be utilized as a part of the ordinary Matlab environment.

Figure 4 shows the components of our application. There are four important components:

- Client user interface
- Google App Engine
- Cloud SQL
- Client browser window

Steps for implementation and deployment of Fuzzy artificial bee colony optimization in the Cloud environment.

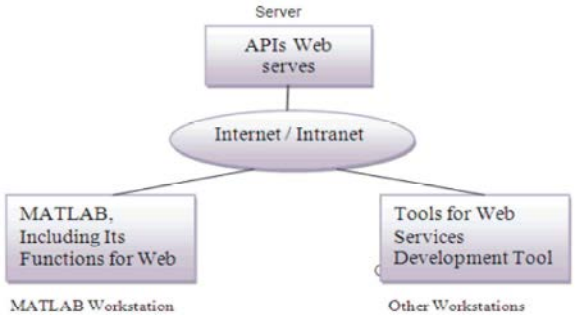


Fig. 3: Web Services in MATLAB

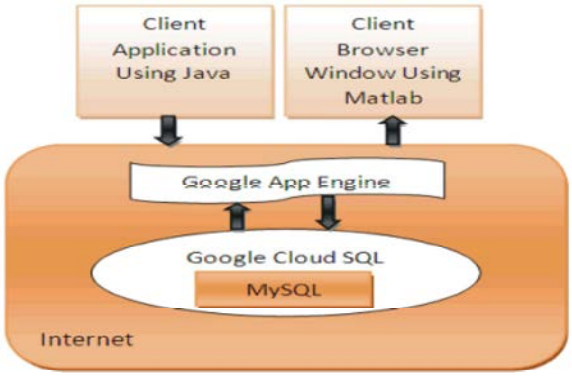


Fig. 4: Architecture of Cloud with Data mining

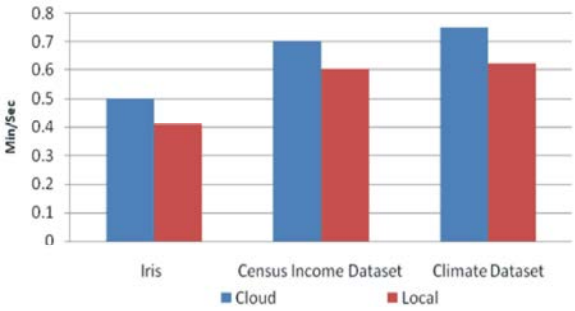


Fig. 5: Comparison of Cloud Storage with three different Cloud dataset

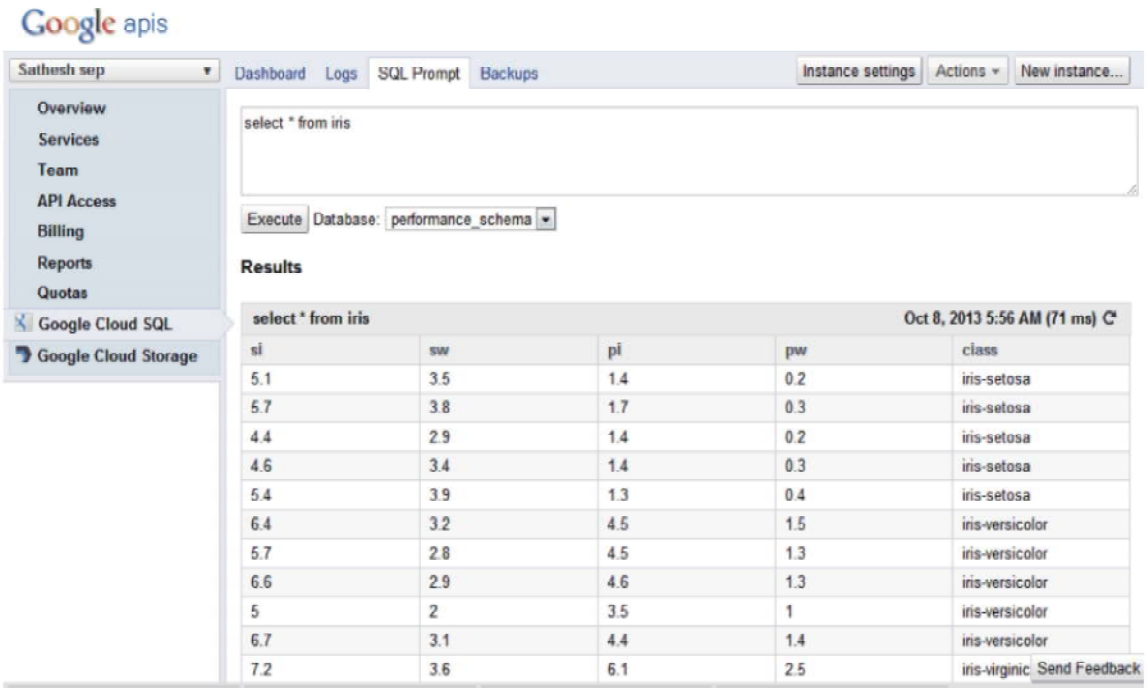
Step 1: Create Database and table in Cloud SQL

Step 2: Take the real-time Dataset from Machine learning repository, 2012 and store it into respective table

Step 3: Write and execute a sample SELECT query and check whether it works well or not in Cloud SQL of Google API's console.

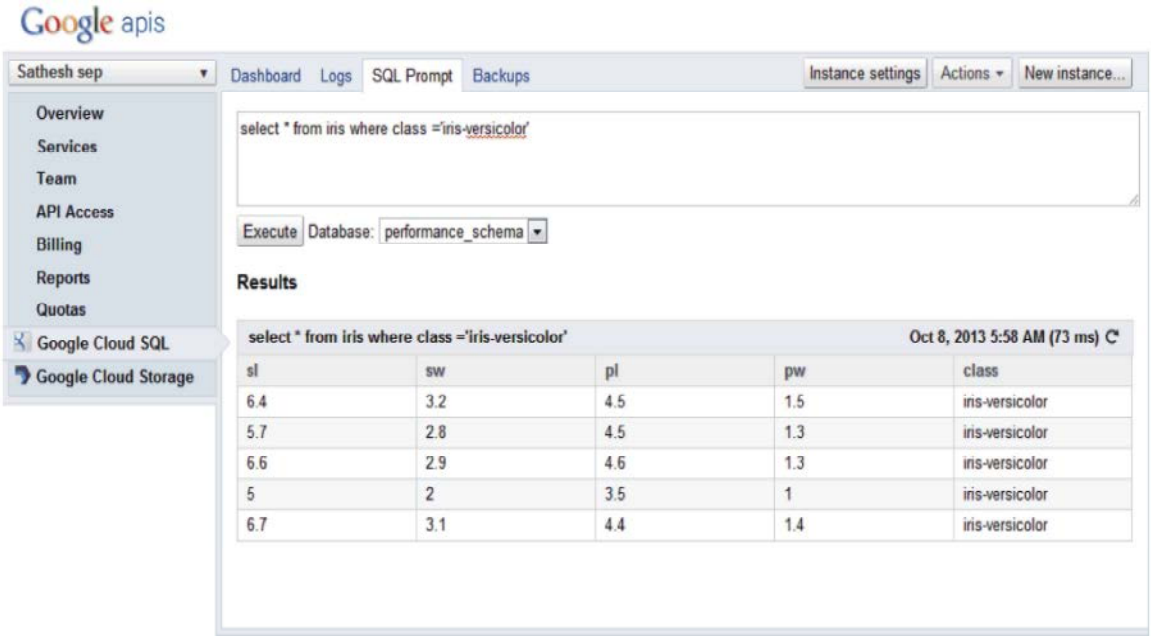
Step 4: Design the User Interface and write code in Matlab for Fuzzy artificial bee colony optimization

Step 5: Debug and deploy the application in Google App Engine Cloud



sl	sw	pl	pw	class
5.1	3.5	1.4	0.2	iris-setosa
4.7	3.8	1.7	0.3	iris-setosa
4.4	2.9	1.4	0.2	iris-setosa
4.6	3.4	1.4	0.3	iris-setosa
5.4	3.9	1.3	0.4	iris-setosa
6.4	3.2	4.5	1.5	iris-versicolor
5.7	2.8	4.5	1.3	iris-versicolor
6.6	2.9	4.6	1.3	iris-versicolor
5	2	3.5	1	iris-versicolor
6.7	3.1	4.4	1.4	iris-versicolor
7.2	3.6	6.1	2.5	iris-virginic

Fig. 6: Storage of IRIS dataset in Cloud SQL



sl	sw	pl	pw	class
6.4	3.2	4.5	1.5	iris-versicolor
5.7	2.8	4.5	1.3	iris-versicolor
6.6	2.9	4.6	1.3	iris-versicolor
5	2	3.5	1	iris-versicolor
6.7	3.1	4.4	1.4	iris-versicolor

Fig. 7: Display records in iris-versicolor

Step 6: Go to the output window of Matlab to view the resultant rules generated.

Figure 5 shows the comparison of using local storage and cloud storage where the later one took more time due to the network traffic but the efficient way of accessing it anywhere at anytime is the added advantage.

Figure 6 shows the storage of Iris dataset in the Google cloud storage. The queries are retrieved from the stored records in the table. It is easy for the multitier architecture where the user is able to access the stored dataset from anywhere.

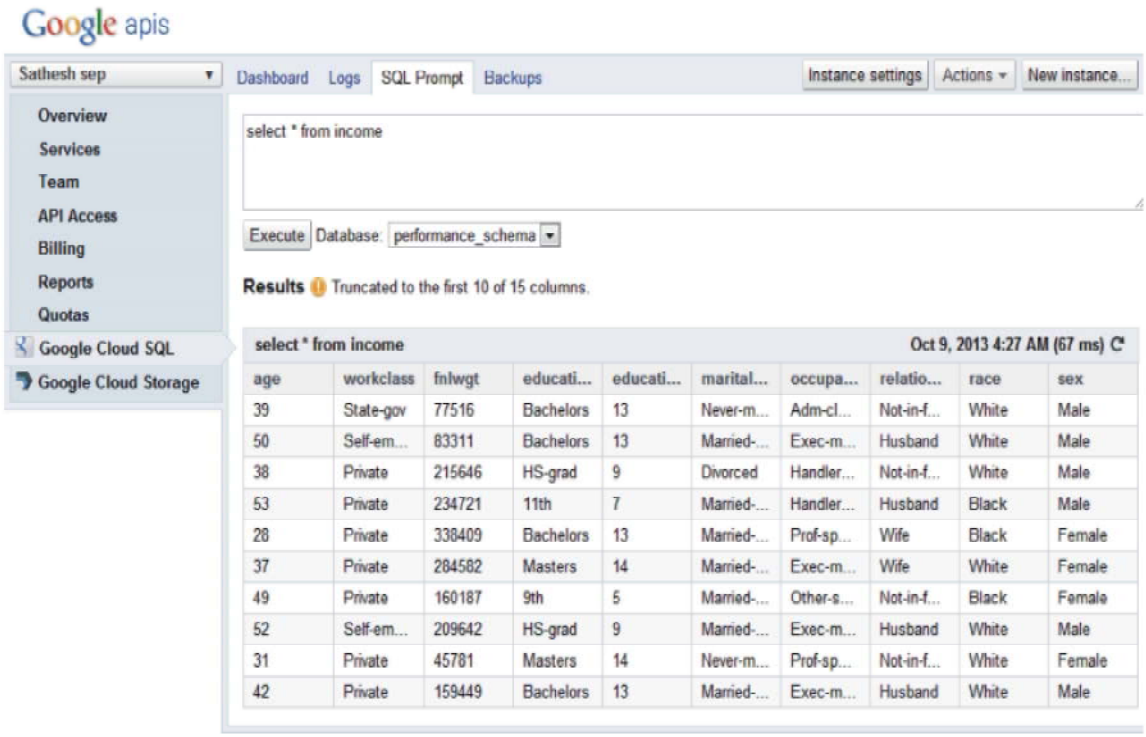


Fig. 8: Display records in Income dataset in Cloud SQL

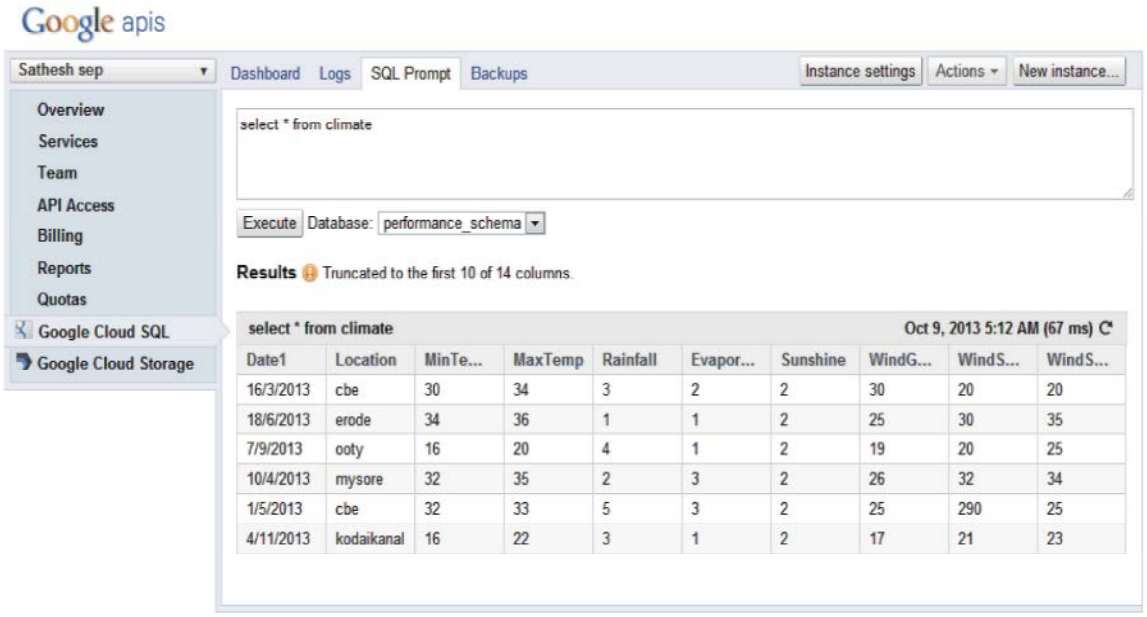


Fig. 9: Displays the records available in the climate table of climate dataset using Google Cloud SQL

RESULTS

As stated the main objective of this research paper is to implement data mining algorithms in Cloud database using Google Cloud SQL, performance result are shown below.

	Configuration
CPU	Intel® Core™2 P7350 @ 2.00GHz
Memory	4.00GB DDR2 SDRAM
Hard-Disk	320GB (5400RPM)
operating system	Windows Vista™ Home Premium

Fig. 10: Configuration of system requirements

Configuration

Dataset Description: For measuring the performance of the proposed FABCO data mining algorithm, an experimental run with different datasets selected from the UCI Machine Learning Repository was conducted. There are 187 data sets currently maintained on the homepage of UCI Machine Learning research group ("UC Irvine Machine Learning Repository,"). The three popular Cloud data sets in the experiment were tested: The Real Time Data set has been taken from open Repository. Iris, Census Income and Climate Dataset. Iris data set consists of a total number of 150 instances. The number of attributes involved in this dataset is 4 with a class Label 3 and Census Income dataset constituted a total number of 48842 instances. The number of attributes involved in this dataset is 14 with a class Label 2 and Climate Dataset constituted a total number of 1000 instances. The number of attributes involved in this dataset is 14 with a class Label 2 as shown in Table 1.

To compare the competence and the accuracy of the proposed work, it was evaluated with the ABC, KNN and PSO. The datasets used in this rule optimization are Iris, Census Income and Climate Dataset. To estimate the reduction process, the average supports of acquired rules on testing data set was calculated. Also the primitive rules on these records were tested. Table 2 shows the results, where the accuracy of fuzzy association rules increased by 12% after 92% reduction of number of rules using FABCO. This shows a good result for reduction of rules of this approach.

From the above Table 3 it is inferred that for the Iris dataset this proposed work reduced the rules to generated optimized classification result efficiently by reducing it to 0.196% compared to other techniques. Table 4 shows that for the Census Income dataset the proposed work reduces the rules to generated optimized classification result efficiently by reducing it to 0.194% compared to other techniques. The above Table 5 shows that for the Climate dataset in which this proposed work reduces the rules to generated optimized classification result efficiently by reducing it to 0.195% compared to other techniques the accuracy of the proposed work performs better after the pruning process. Generally, from the overall performances, the current research study outperforms other techniques after pruning and its utilization is strongly recommending [17].

Performance Comparison based on Precision, Recall, F-measure and Time taken:

Table 1: Description of Three Different Cloud Dataset

Dataset	Features	Instances	Class
Iris Dataset	4	150	3
Census Income Dataset	14	48842	2
Climate Dataset	14	1000	2

Table 2: Performance Accuracy of classification

	KNN	PCO	ABC	FABCO
Iris Dataset	0.76542	0.85491	0.89773	0.98517
Census Income Dataset	0.81068	0.83192	0.91974	0.97693
Climate Dataset	0.81021	0.87315	0.95921	0.99003

Table 3: Accuracy percentage of Iris Dataset

Techniques	Primitive Rules	Reduced Rules
PSO	0.48	0.320
ABC	0.39	0.297
FABCO	0.20	0.196

Table 4: Accuracy percentage of Census Income Dataset

Techniques	Primitive Rules	Reduced Rules
PSO	0.35	0.318
ABC	0.30	0.261
FABCO	0.21	0.194

Table 5: Accuracy percentage of Climate dataset

Techniques	Primitive Rules	Reduced Rules
PSO	0.41	0.385
ABC	0.35	0.297
FABCO	0.20	0.195

- Precision is the probability that a (randomly selected) retrieved rule is relevant.

$$\text{Precision} = \frac{tp}{tp + fp}$$

- Recall is the probability that a (randomly selected) relevant rule is retrieved in a search.

$$\text{Recall} = \frac{tp}{tp + fn}$$

- F-Measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2.\ \frac{precision.recall}{precision + recall}$$

Chart 11, 12 and 13 shows the performance comparison of proposed FABCO with existing approaches for classifying the three different dataset namely Iris, Census income data and Climate dataset. The result shows that the proposed approach has highest precision,

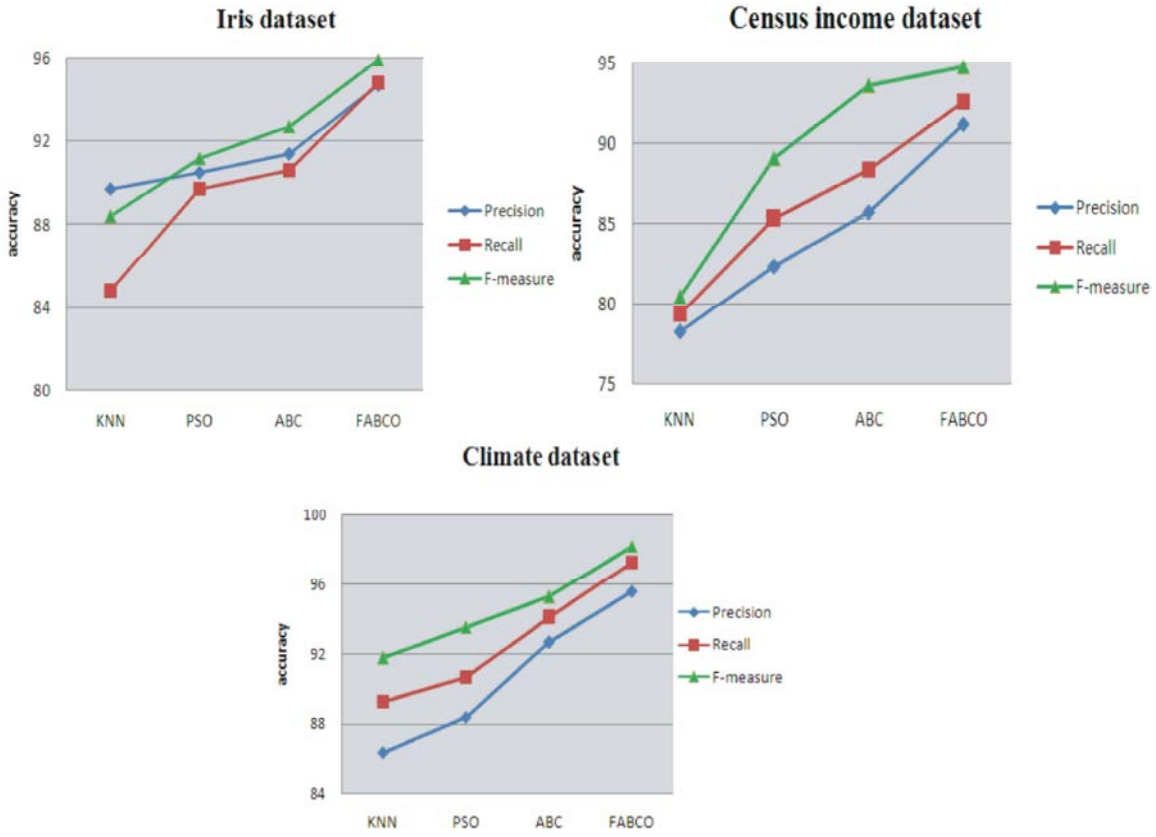


Fig. 11, 12, 13 shows the performance comparison of proposed approach with the already three other existing approaches

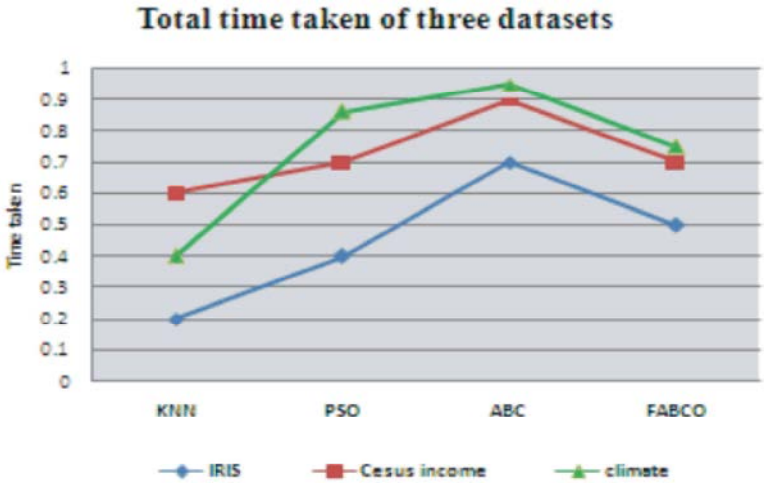


Fig. 14: Total time taken of three different datasets

recall and F-measure value. Figure 14 shows that lowest time was taken by K-NN technique but its performance was the worst compared to other approaches.

The chart shows that the proposed work recorded highest precision value, recall and f-measure compared to the existing approaches with the percentage of 94.7, 94.8 and 95.9 in Iris dataset respectively.

CONCLUSION

In this research work, Fuzzy artificial bee colony optimization was proposed to fine tune the generated Fuzzy association rules in order to improve the classification accuracy by analyzing three different real world dataset to demonstrate mining large Databases

Cloud computing provides solution for storing large database with less cost. Thus, use of the proposed algorithm in cloud environment with higher data storage cost-effectively is strongly recommended.

REFERENCES

1. Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg and Ivona Brandic, 2008. Cloud computing and emerging IT platforms: Vision, hype and reality for delivering computing as the 5th utility, Future Generation Computer Systems, Elsevier.
2. Jiguang Wan, Zhuo Liu and Peng Wang, 2010. Data Mining of Mass Storage Based on Cloud Computing. Grid and Cooperative Computing International Conference, pp: 426-431.
3. Haishan Chen, Lu Huang and Xiaodan Zhu, 2012. A survey of mass data mining based on cloud-computing. Anti-Counterfeiting, Security and Identification International Conference, pp: 1-4.
4. Min Zhang, 2011. The Strategy of Mining Association Rule Based on Cloud Computing. Business, pp: 475-478.
5. Yongzheng Lin, 2012. Study of Layers Construct for Data Mining Platform Based on Cloud Computing, Network Computing and Information Security Communications in Computer and Information Science, Springer, 345: 106-112.
6. Mahendiran, A., N. Saravanan, N. Venkata Subramanian and N. Sairam, 2012. Implementation of K-Means Clustering in Cloud Computing Environment, Research Journal of Applied Sciences, Engineering and Technology, 4(10): 1391-1394.
7. Michael Shindler, Alex Wong and and Adam Meyerson, 2011. Fast and Accurate k-means For Large Datasets, Neural Information Processing System Foundation, Conference.
8. Ashok, P., G.M. Kadhar Nawaz, E. Elayaraja and V. Vadivel, Improved Performance of Unsupervised Method by Renovated K-Means.
9. Chang, Y.S., C.T. Yang and Y.C. Luo, 2011. An ontology based agent generation for information retrieval on cloud environment, Journal of Universal Computer Science, 17(8): 1160-1135.
10. Jain, V., 2012. Information retrieval through multi-agent system with data mining in cloud computing, IJCTA, 3(1): 62-66.
11. Zeba Qureshi, Jaya Bansal and Sanjay Bansal, 2013. A Survey on Association Rule Mining in Cloud Computing, International Journal of Emerging Technology and Advanced Engineering, 3(4): 318-321.
12. Yang, S.Y., D.L. Lee, K.Y. Chen and C.L. Hsu, 2011. Energy-saving information multiagent system with web services for cloud computing, SUComS, 2011. CCIS, 223: 222-233.
13. Yu Mon Zaw and Nay Min Tun, 2013. Web Services Based Information Retrieval Agent System for Cloud Computing, International Journal of Computer Applications Technology and Research, 2(1): 67-71.
14. Jiawei Han and Micheline Kamber, 2006. Data Mining Concepts and Techniques?, ISBN: 978-1-55860-901-3.
15. Kalyani Mali and Samayita Bhattacharya, 2013. Fingerprint Database Handling Using Cloud Computing With Added Data Mining and Soft Computing Features, International Journal of Emerging Technology and Advanced Engineering, 3(2): 610-617.
16. Dion, H., 2008. Comparing Amazon's and Google's Platform as-a-Service (PaaS) Offerings. Reterived from: http://www.zdnet.com/blog/google/ theproblem- with-google-apps-engine/1002.
17. Lai, C.F., J.H. Chang, C.C. Hu, Y.M. Huang and H.C. Chao, 2011. A cloud-based program recommendation system for digital TV platforms, Journal Future Generation Computer Systems, 27(6). Retrieved October 25, 2011 from http://dl.acm.org/ citation.cfm?id=1967928.