ENCODED POLYMORPHIC ASPECT OF CLUSTERING

Mrs. A.JACKULIN SAM GINI [1], A.SRINIVASAN [2], S.VIJAYAKUMAR [3]

[1] Assistant Professor, Dept. of IT, Jeppiaar SRR Engineering College, Chennai, Tamil Nadu
[2,3]B.TECH., Dept. of IT, Jeppiaar SRR Engineering College, Chennai, Tamil Nadu

***

Abstract - *Data storage are constantly growing and maintaining the large data it leads to more complex. We face the difficulty of handling the data and storage problems. The data can be getting from various sources and analysis in a different views are referred to as multi-view data. So we propose the Machine learning technologies have been investigated for the scope of dealing with multi-view data. This paper focuses on one of the unsupervised learning techniques, namely, Clustering. It means that similar objects are grouped into the same cluster, and dissimilar objects are divided into different cluster. Compared to single-view clustering, multi view clustering normally can access to more characteristics and structural information hidden in the data, and intuitively can exploit richer properties of data to improve the clustering performance.*

1. INTRODUCTION

Machine learning is to predict the future from past data. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using java. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be divided into three categories such as supervised learning, unsupervised learning and reinforcement learning. In Supervised learning ,the machine is provided with a new set of data so that supervised learning analyzing the dataset and produces the correct ouput.In Unsupervised learning we don't need any labels. It allowing the algorithm to access the dataset without any guidance. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance. Data scientists use many different kinds of machine learning algorithms to discover patterns in java that lead to actionable insights. Clustering means dividing the dataset into a number of groups such that similar datasets are clustered in the same groups and more similar to other datasets also in the same group and dissimilar dataset in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. There are no criteria for a good clustering. Clustering based on the user requirement and grouping of data must satisfying the user needs.

1.1 What is Unsupervised Machine Learning?

Unsupervised learning have a capability of self learning algorithm without any associated trained datasets, this algorithm easily determine the data patterns on its own way. This type of algorithm leads to restructuring the data into something different, such as new features that may represent in a new group of uncorrelated values. Unsupervised learning is the training of machine using information there is no classification or labeling of any data its allowing the algorithm to act on that dataset without guidance. Here the task for grouping the unsorted information based on data similarities, analyzing in depth of data patterns and differences without any previous trained datasets. Unlike supervised learning, no trainer is provided that means no previously trained datasets will be given to the machine. So, machine is restricted to find the in depth unlabeled data structure by its self. Unsupervised learning have two categories of algorithms such as Clustering and Association. Clustering method is where you want to discover the grouping of polymorphic data, such as grouping students based on their score. Wherever you want to discover the rules and describing the big datasets then we can use association algorithm.

1.2 What is BMVC ?

BMVC means Binary Multi-View Clustering algorithm, which can analyzing the multi-view image data and easily scalable of big datasets. To achieve this goal, we use two types of BMVC methods. Collaborative discrete data representation method and Clustering the structure of binary data, BMVC collaboratively encoding the image in multiple way descriptors into a common binary code space by considering their complementary data; For collaborative binary representations of clustering are done by binary matrix factorization method, such that the cluster structures are optimized in the Hamming space by pure, and fast bit-operations. K-means algorithm has severely unaffordable computational time and storage requirement in real-world applications with large data and a big number of clusters. Recently, multi-view clustering by exploiting heterogeneous features of data has attracted considerable attention. For efficiency, the code balance constraints are imposed on both binary data representations and cluster centroids. These algorithms focus on speed and scalability.

2. MODULES DESCRIPTION

2.1 UPLOAD THE DATASETS

The information is transferred from the establishment which is situated in various areas transfers separately. All the transferred information is gotten by the director and afterward assembled in one spot. Candidates upload the basic details. The exam conducted under four categories are listening, reading, writing, speaking. After completion of exam the result will be upload by admin that four categories in listening, reading, writing, speaking. The score upload in between the particular range from 8 to 40.

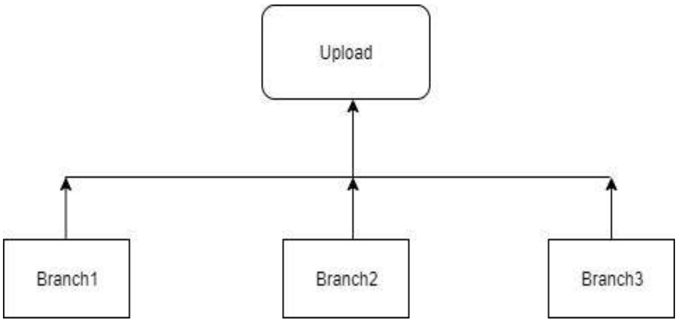


Fig-1 Upload the Datasets

2.2 BINARY CONVERSION

Before datasets are being uploaded it should convert into binary. For the conversion, we use the binary matrix factorization method. These algorithms focus on speed and scalability they work with binary factors combined with bit-wise operations and a few auxiliary integer ones. For binary conversion we use in-built package java. lang. Inside the java.lang package we calling the parseByte() method from that package for conversion. All the upload dataset are converted into binary code. So, the memory space of every data is less. For safe and security purpose all the data should be encoded. For encoding the data we use Base 64 method in java. Base 64 is an encoding scheme that converts binary data into text format so that encoded textual data can be easily transported over network uncorrupted and without any data loss. Binary File have less memory allocation and quick access based on user query. Its save more time for users or applicant.

File
Binary Code
(eg.1010,1100)

Fig-2 Binary Conversion

2.3 CLUSTERING THE DATA

In unsupervised learning give the training to the machine using data and there is no classification or labeled data and allowing the algorithm to act on that data without guidance. Machine is restricted to find the in depth data structure from unlabeled data by its self. There is an task for machine to group the unsorted and unordered information depending on their data similarities, data patterns and differences without any past training of datasets. It is used as a process to find meaningful data structure, underlying pattern processes, generative the additional features, and groupings of data. Categorization of the data sets into a number of groups such that similar data are in the same groups and dissimilar data sets are in other groups.

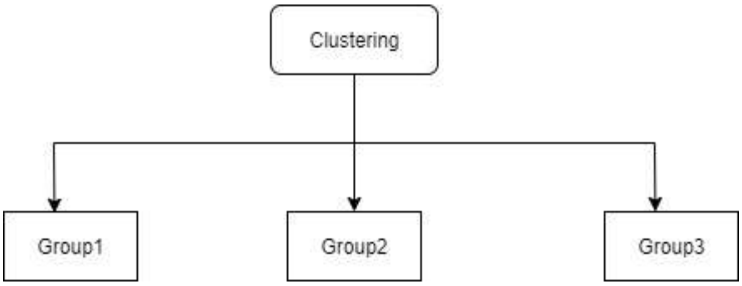


Fig-3 Clustering the data

2.4 RETRIEVING THE DATA

The data retrieval is the process of identifying and extracting data from a database. Based on a query provided by the user or application the data should be retrieve from the database. In our project all the data are stored in binary encoded format. So, First we decoded the data and then view the data based on query. Base 64 is an decoding scheme method that converts text data into binary data format so that decoded binary data can be easily transported over network. The candidate data will be segregated, after the result uploaded by the admin. The segregated data can be view under four categories such as Listening, Reading, Writing and Speaking. The recruiter choose atleast any one of the field or all the four fields for segregating the candidate results. This segregation method is very useful for recruiter to shortlist or select the eligible candidate easily. For safe and security purpose this encryption and decryption method is very useful.

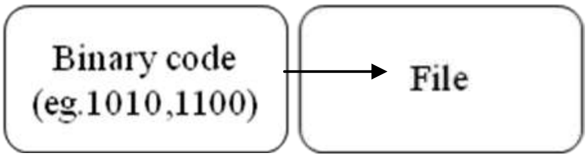


Fig-4 Retrieving the Data

3. SYSTEM TECHNIQUES

The Software Development Lifecycle is very helpful for the completion of software development and implementation. This process also helpful for developing a complex software applications. SDLC is a process followed by the organization for developing a software project, only by software organization. SDLC contains detail plan for developing, maintaining, testing, updating, replacing and altering or enhancing the specific software. This life cycle for developing a software defines a various ideas for improving the quality of software product and the overall development of project. The major concepts of this document is to present a complete process of the Web application system. This document is useful for both the stakeholders and the developers.

3.1 SYSTEM ARCHITECTURE

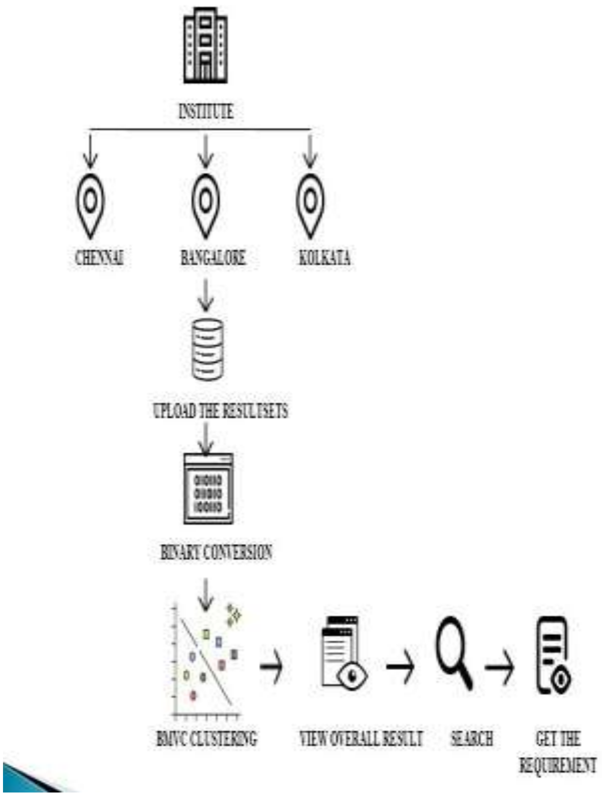


Fig-5:Architecture Diagram

4. HARDWARE REQUIREMENTS

- **Processor** : i3/i4
- **System** : Pentium Dual-Core
- **Hard disk** : 120 GB and Above
- **RAM** : 2 GB and Above

4.1 SOFTWARE REQUIREMENTS

- **Operating System** : Windows
- **Front end** : Core Java,CSS, JSP,Servlet
- **Web application** : J2EE,Hibernate
- **Back end** : MySQL 5.1
- **Tool** : Eclipse

4.2 FUNCTIONAL REQUIREMENTS

Functional requirement defines as a specification of behavior between output and input in system and software engineering. Based on that requirement engineering, functional requirements describing the particular results of a system. Some of the more typical functional requirements include business rules, authentication, authorization, legal requirements etc;

4.3 NON FUNCTIONAL REQUIREMENTS

Non-functional requirements define how the system works, but in functional requirements.It describe what the system should do. Non-functional requirements mainly focused on the quality attributes of a system. They specify criteria that judge the operation of a system, rather than specific behavior such as performance, scalability, availability, maintainability, reliability, data integrity, security etc;

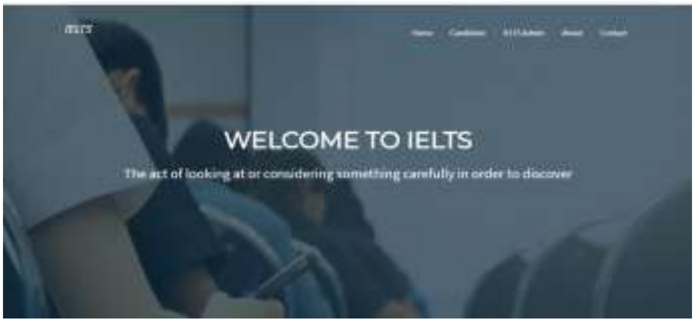
5. INPUT

- The user must create the account for login. All the user details have been stored the data in our database for future purpose.
- The user view the exam schedule and book the date for attending the test and upload the location and timing also.
- The admin verify the candidate details and then approve/deny based on their application.
- The admin upload the candidate result.

5.1 OUTPUT

- The Application owner can show the user details and it can validate some sensitive information details.
- The user can also get their admit card, acknowledgement for enrollment and result from the admin side.
- In this section recruiters/user can give the input for segregating the result and get the output from the admin side based on the user query.

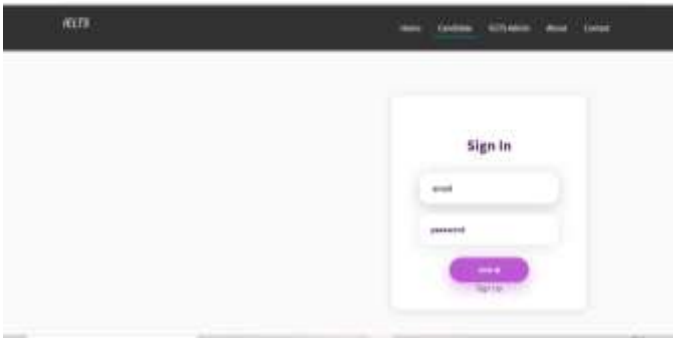
6. SNAPSHOT



WELCOME TO IELTS

The act of looking at or considering something carefully in order to discover

Fig-6: Home page

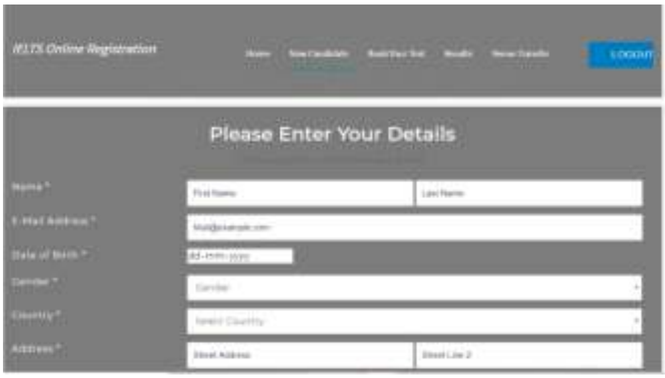


Sign In

email

password

Fig-7: Login page



Please Enter Your Details

Name * First Name Last Name

E-Mail Address * Mail@example.com

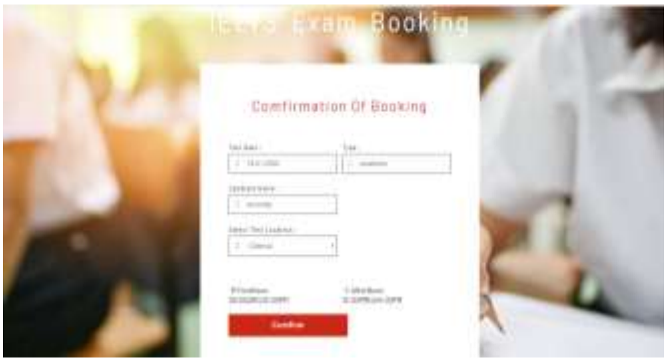
Date of Birth * dd-mm-yyyy

Gender * Gender

Country * Select Country

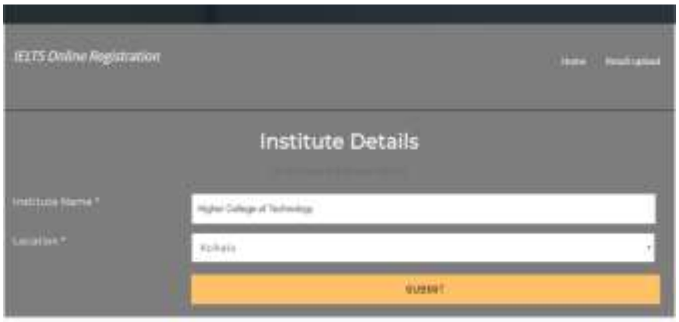
Address * Street Address Street Line 2

Fig-8: Candidate Enrollement



Comfirmation Of Booking

Fig-9: Booking for Test

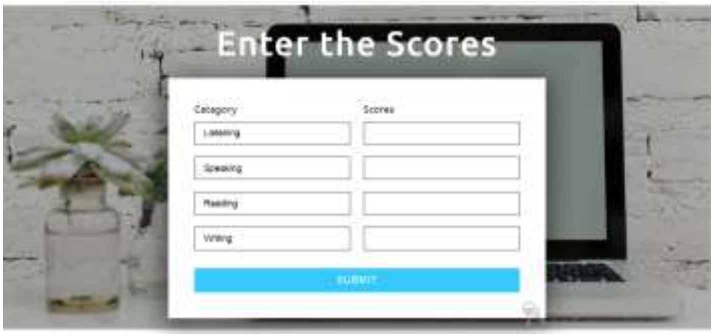


Institute Details

Institute Name * Higher College of Technology

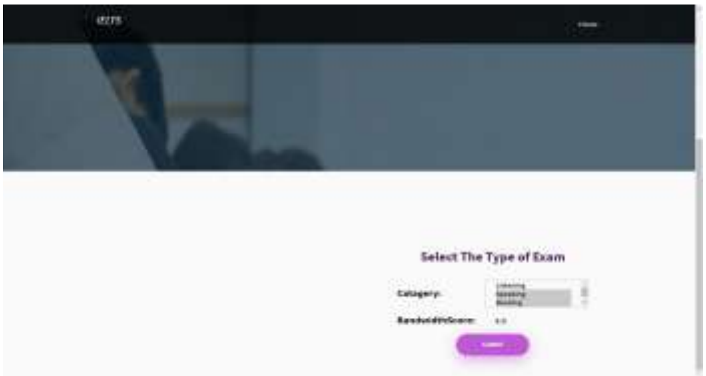
Location * Kolkata

Fig-10: Institute Details



Category	Scores
Listening	
Speaking	
Reading	
Writing	

Fig-11: Upload the result



Select The Type of Exam

Catagery: Listening Speaking Reading

BandwidthScore: [illegible]

Fig-12: Result Segregation

LIST OF CANDIDATE

ApplicationID	CandidateName
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Fig-13: Segregated Result

7. ADVANTAGES

The scope of this project stores the data is secured and fast in performance and enhances storage capability.Moreover, we use binary code conversion to reduce the more memory consumption. BMVC techniques to increase the fast optimization of the data viewing and analysis.When we use BMVC, it give exact and accurate of information more than the other clustering methods. Encoding of data is very helpfu for safe and security purpose.

8. APPLICATION

The company recruiters want to select or shorlisted the candidate based on there score. Admin upload the candidate result under four category such as Listening, Speaking, Reading and Writing. After the completion of the uploading the result clustering and segregation process to be done. The user can also view the result login through with there respective use account. Finally, the recruiters segregate the candidate based on there needs.

9. FUTURE ENHANCEMENT

What's more, presenting the encoded bunching procedure and improvement strategies were joint together and increment the compelling calculation in the bigger datasets. Later on, further encourage the presentation, we intend to explore the administered and profound augmentation of encoding in-see portrayal by the AI system. In the AI method, the profound learning framework is utilized for directed learning of the system. Be that as it may, it executes a profound neural system to use the accessible name.

10. CONCLUSION

In this paper, a principled Binary Multi-View Clustering (BMVC) method, was proposed for solving the challenging problem of multi-view clustering on large-scale image data. In BMVC, the collaborative discrete representations and binary cluster structures were jointly learned, which could effectively integrate the collaborative information from multiple views. Moreover, an effective alternating optimization algorithm with guaranteed convergence was proposed to ensure the high-quality binary solutions.

11. REFERENCES

[1] A Survey on Learning to Hash, J. Wang, T. Zhang,02 May 2017, https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/LTHSurvey.pdf

[2] Binary Multi View Clustering, Zheng Zhang, Ling Shao ,July 2019,https://ieeexplore.ieee.org/document/8387526

[3] Compressed K-Means for Large-Scale Clustering, X. Chen, W.Liu.,2017, https://www.semanticscholar.org/paper/Compressed-K-Means-for-Large-Scale-Clustering-Shen-Liu/93c03ff9421c49b74c234d6486e8884b6d744b54

[4] Fast K-Means with Accurate Bounds,J. Newling, F. Fleuret,Sep 2016, https://arxiv.org/pdf/1602.02514.pdf

[5] Learning Short Binary Codes for Large-scale Image Retrieval, Li Liu, Mengyang Yu.,11 January 2017, https://www.freeprojectsforall.com/wp-content/uploads/2018/09/Learning-Short-Binary-Codes-for-Large-scale-Image-Retrieval.pdf

[6] Multiview Alignment Hashing for Efficient Image Search, Mengyang Yu,12 January 2015, https://ieeexplore.ieee.org/abstract/document/7006770

[7] Multi-View Clustering via Joint Non-negative Matrix Factorization,JialuLiu , Chi Wang , Jing Gao, and Jiawei Han,May-2013, https://www.researchgate.net/publication/279953559 Multi_View_Clustering_via_Joint_Nonnegative_Matrix_Factorization

[8] Unsupervised Deep Hashing with Similarity-Adaptive and Discrete Optimization,Fumin Shen, Yan Xu,Yang Yang,05 January-2018, https://ieeexplore.ieee.org/abstract/document/8247210

.