

# An advanced ilrepsd technique for bridging the competency and cognitive skills of students in higher education

Joy Christy. A\*

Department of Computer Science Engineering, SASTRA University, Thanjavur – 613 401

\*Corresponding author E-mail: [joychristy@cse.sastra.edu](mailto:joychristy@cse.sastra.edu)

## Abstract

Data mining refers to the extraction of meaningful knowledge from large data sources as it may contain hidden potential facts. In general the analysis of data mining can either be predictive or descriptive. Predictive analysis of data mining interprets the inference of the existing results so as to identify the future outputs and the descriptive analysis of data mining interprets the intrinsic characteristics or nature of the data. Clustering is one of the descriptive analysis techniques of data mining which groups the objects of similar types in such a way that objects in a cluster are closer to each other than the objects of other clusters. K-means is the most popular and widely used clustering algorithm that starts by selecting the k-random initial centroids as equal to number of clusters given by the user. It then computes the distance between initial centroids with the remaining data objects and groups the data objects into the cluster centroids with minimum distance. This process is repeated until there is no change in the cluster centroids or cluster members. But, still k-means has been suffered from several issues such as optimum number of k, random initial centroids, unknown number of iterations, global optimum solutions of clusters and more importantly the creation of meaningful clusters when dealing with the analysis of datasets from various domains. The accuracy involved with clustering should never be compromised. Thus, in this paper, a novel classification via clustering algorithm called Iterative Linear Regression Clustering with Percentage Split Distribution (ILRCPSD) is introduced as an alternate solution to the problems encountered in traditional clustering algorithms. The proposed algorithm is examined over an educational dataset to identify the hidden group of students having similar cognitive and competency skills. The performance of the proposed algorithm is well-compared with the accuracy of the traditional k-means clustering in terms of building meaningful clusters and to prove its real time usefulness.

**Keywords:** Pearson Correlation, EDM, Clustering, PSD, Silhouette Distance

## 1. Introduction

The term EDM refers to the transformation of raw educational data into knowledgeable facts that could potentially helpful for the educators to afford the best educational practices. EDM is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context [1] (Romero et al, 2010). The techniques of EDM have been extensively applied in numerous educational applications and they offer realistic output in all educational aspects. EDM strongly contributes to the betterment of imparting quality in higher education by screening several applications in new dimensions. At present, educational data mining is paid great attention by most of the researchers, since it helps in improving the current educational systems from performance prediction to retention. Though there have been plentiful efforts made in EDM, the analysis on bridging the gap between the cognitive and competency skills of the students in terms of their employability still seems to be diminutive as the students graduating in Indian colleges run the risk of being unemployed. Barely 3.84 per cent of students are geared up for employability in start-up companies. In general, the learning abilities of the students in higher education can be classified as cognitive, affective and psychomotor domains. Cognitive learning domain is connected with brain that encompasses the development of knowledge with consistent recall

and recognition of academic materials. Affective learning domain is connected with the heart which in turn represents the attitude and skills of the students towards the academic materials [2] (Allen et al, 2010). Finally the psychomotor learning domain is connected with the hands, which indicates the practical knowledge of the students on the academic materials [3] (Bolin et al, 2005; Ferris et al, 2005). Sometimes the students who have great cognitive learning domain may not get placed in the campus interviews whereas the students who have great affective and psychomotor learning domains may get selected as their competency skills may be higher than the students with higher cognitive skills. Here, the competency skills refer to both affective and psychomotor learning domains. The identification of students' performance prediction in campus interviews helps them be focused on their career development. The students' knowledge on the learned materials alone does not support them to get through the interview; rather the interviewers are very much keen on the selection of students with high cognitive and competency (affective and psychomotor) skills. Hence, the analysis on students' learning domain is inevitable for the educators, which is still an issue in many institutions. The research problems that motivate the thesis are:

- Students' success in campus interviews are still diminutive
- There is no technology available to identify the competency and cognitive skills of students in higher education
- There is no effort made to group the similar types of students with the same skills

- There are no special programmes conducted to improve the identified lacking skills of the students

The purpose of identification of students' learning domain helps to group the students with similar learning domain so as to impose adequate training programmes to improve their skills that they are lacking in. Identification of students' learning domain and grouping of students with the same learning domain come under the scope of clustering techniques as it has the ability to group the data with similar characteristics. Clustering methods stay as one of the predominant machine learning techniques for grouping of similar instances. Hence, the grouping of students with the same learning domain may also be achieved by clustering methods. The aim of this research paper is to propose an advanced clustering technique that bridges the competency and cognitive skills of the students in higher education.

## 2. Review of Literature

Clustering is one of the frequently used data mining techniques which is used to exhibit the similarity in data objects over large datasets. Clustering techniques group the homogenous data objects (records or instances) in such a way that objects in a cluster are closer to each other than the objects of other clusters based on certain similarity measures. Clustering techniques analyze the divergence of data objects so as to analyze the future initiations. The types of clustering are majorly categorized into connectivity, centroid, distribution and density based techniques. Each type of clustering techniques has its own merits and demerits in terms of building quality clusters. This section reviews the existing literature on types of clustering used in practice and the recent works carried out to enhance the demerits of the traditional clustering algorithms.

K-means is known as the best understood model as ever, that tends to build clusters of comparable multidimensional spatial extent. The method first selects 'k' random initial centroids and computes the distance between the centroids and data objects. The centroids are originating clusters by grouping the data objects with minimum distances. The centroids are recomputed and the cluster members are reassigned till there is no change in the cluster members or cluster centroids. This model is iterative and the centroids of the clusters may or may not be a member of the dataset. The model requires the number of clusters to be specified in advance. The time complexity of K-means clustering is computed as  $(n+K+t)$ , where 'k' specifies the number of clusters, 'n' denotes the total number of data objects, and 't' represents the number of iterations [4] (Yadav et al, 2013). Moreover, k-means algorithm is the most suitable one for grouping numerical data and can be modified to adapt to categorical data.

Wu et al (2015) [5] have stated that the classic centroid models are sensitive to the selection of initial centroids and often converge with locally optimized solutions. Hence, the authors have presented a Hybrid Fuzzy K-Harmonic Means (HFKHM) clustering that combines an Improved Possibilistic C-Means (IPCM) with KHarmonic Means (KHM). IPCM incorporates fuzzy approaches on to PCM's objective function so as to form appropriate clusters and also to minimize the noise sensitivity. KHM employs the harmonic means of distances between the centres and data points as components to the objective function.

Prabha et al, (2014) [6] have proclaimed that the centroid models have noise and selection of initial clusters sensitivity and often terminate at a local optimum. The authors have collaborated the concept of K-means with particle swarm optimization so as to present solution for the convergence of global optimum. The authors have employed min-max normalization as a preliminary step before clustering for yielding better cluster results. PSO works on the basis of swarming the number of individual particles with respect to their velocities. The swarms are then adjusted by

comparing the local and global position of data through a number of iterations.

Chawla et al. (2013) [7] have indicated that the centroid models are extremely sensitive to outliers and presented a naïve approach to discover the outliers. The authors have proposed K-means algorithm that discovers 'k' clusters and 'l' outliers in each iteration. The distance between the data points and the centroid is decreasingly ranked and the top 'l' data points are stored into a list which is then removed from the list to be formed as a separate cluster. The distance between members of the new cluster is iteratively computed and assigned to the nearest cluster until the members of the clusters are stabilized.

Anand et al, (2015) [8] have denoted that the centroid models are computationally complex and build clusters with poor quality. Hence, the authors have developed a modified k-means algorithm that focuses on a few issues right from the selection of initial centroid till the reduction of computational cost which compares only the boundary elements of each cluster. A fuzzy membership function is also performed in each step to analyze the stability of the centroids. To minimize the unwanted moves of data objects, the method stores the current centroid and compares it with the old centroid and does not make changes when the current centroid is less than the old one.

The problems identified with centroid model are:

- Bad initialization
- Often terminates at a local optimum
- The cluster has spherical shapes
- Sensitivity to noise
- No criteria for selecting the number of clusters
- Detection of outliers
- Creation of empty clusters
- High computational complexity

Clustering is one of the prevalent data analytics techniques of data mining that interpret a large dataset, as the manual classification is more complex and time consuming. Though k-means is the most simple and widely used clustering algorithm, it suffers from certain important issues. Hence, the research on the improvisation of k-means clustering techniques has always been considered as a thrust area. Moreover, the invention of advanced clustering techniques has also become mandatory for clustering students with the same learning ability.

Classification algorithms take the advantage of the labeled classifier for the prediction of classes of data objects which is called as supervised learning of data. Classification algorithms build the knowledge prediction models by training the data objects with labeled classifier for predicting the classes for test data objects. Hence, the contribution of labeled classifier is significant in predicting the classes of the dataset. The clustering algorithm proposed in this chapter also takes the advantage of labeled classifier for correctly clustering the data objects based on the classes so as to improve the accuracy of clustering with the following objectives:

- To propose a clustering algorithm that enhances the clustering performance in terms of accuracy

To generate meaningful clusters using labeled classifier.

## 3. Methodology

This paper proposes a novel Iterative Linear Regression Clustering algorithm with Percentage Split Distribution (ILRCPSD) method with the objective of clustering the instances of similar items through classification via clustering approach. The algorithm iteratively transforms the attributes into a linear form using multiple linear regression algorithm for attaining the highest correlation of attributes with the class variable. The coefficient of attributes retrieved in each iteration is multiplied with the instance values which

are then grouped using Percentage Split Distribution (PSD) method. PSD splits the data objects according to its distribution in regression line. The proposed ILCRPSD algorithm executes all the functional requirements of data mining from feature selection to performance validation. The framework of the proposed methodology is depicted in Fig 1.

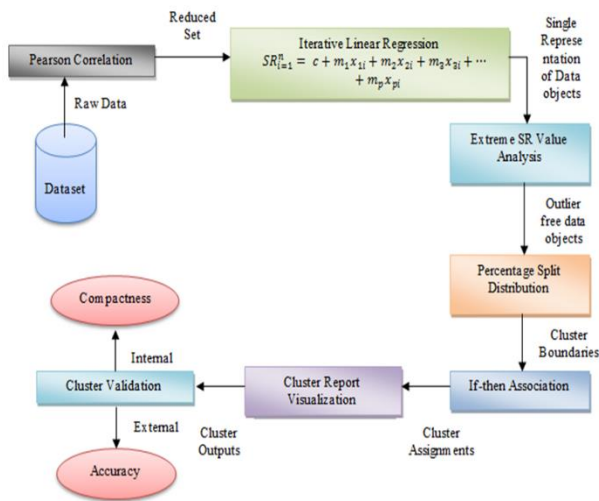


Fig.1: Framework of Proposed ILCRPSD

### 3.1. Pearson Correlation

In ILCRPSD, Pearson correlation method is used for feature selection to extract the best set of attributes that influences the class variable. The method investigates the measure of linear relationship between the dependent and independent variables with three feedbacks such as positive, negative and no correlation. In this research, the features that are positively and negatively correlated with class variable are selected as the linear correlating variables improve the strength of the regression line. The Pearson correlation method is computed using the formula given in Equation 1.

$$PCC = \frac{\sum(I-I')(D-D')}{\sqrt{\sum(I-I')^2 \cdot \sum(D-D')^2}} \quad \dots \quad (1)$$

Where

$I$  is the independent variable

$D$  is the dependent variable

$I'$  is the mean of  $I$

$D'$  is the mean of  $D$

### 3.2. Iterative Linear Regression

To increase the correlation of independent and dependent variables, the features that are selected from the Pearson correlation are executed upon linear regression so as to fit into a regression line. A linear regression model that is formed with single dependent (scalar) and independent (explanatory) variable is called simple linear regression model which is denoted in Equation 2.

$$y = mx + b \quad \dots \quad (2)$$

Where 'm' is a slope and 'b' is an intercept on y-axis. The computation of slope 'm' is represented in Equation 3 and line intercept on y-axis 'b' is denoted in Equation 4.

$$m = c \times \frac{S_e}{S_s} \quad \dots \quad (3)$$

' $S_e$ ' is the standard deviation of explanatory variable, ' $S_s$ ' is the standard deviation of scalar variable and 'c' is the correlation value of explanatory and scalar variables.

$$c = \text{mean}_y - m \times \text{mean}_x \quad \dots \quad (4)$$

A linear regression that is formed with single dependent and multiple independent variables is called a multiple linear regression which can be denoted with a notation presented in Equation 5 for any object 'i'.

$$y_i = c + m_1x_{1i} + m_2x_{2i} + m_3x_{3i} + \dots + m_px_{pi} \quad \dots \quad (5)$$

Where 'c' is the line intercept of y-axis, 'm' is slope of line. In this work, the regression model is iteratively executed to increase the linear relationship of independent and dependent variables closure to the maximum or until it reaches the stable regression line. The accumulation of processed attributes of the notations forms a Single Representation (SR) of data objects.

### 3.3. Extreme SR Value Analysis

The presence of outliers in ILCRPSD leads to an abnormal SR distribution of data objects which could bias the cluster splits. Hence, an Extreme SR Value Analysis method with single representation of data objects is implemented to identify the outlier data objects. This method is employed to detect the existence of outliers in the SR of data objects as the outliers may influence the results of percentage split distribution. The median R2 is obtained by finding the median of SR values. The SR values are arranged to the lowest to highest order. The median R2 is the data object that splits the data into two halves. It is easy to find the median of the dataset if holds odd number of data objects. If there are even number of data objects two middle values of the data objects are averaged to find the median. From the median R2 the lower quartile median R1 is computed by calculating the median of the values below R2 and the upper quartile median R3 is computed by calculating the median of values above R2. The inter-quartile median RQ is calculated by subtracting the upper quartile median with the lower quartile median, (i.e) R3-R1. The outlier data objects are identified by assessing whether or not they fall within a set of numerical boundaries called inner and outer fences. The lower inner fence is calculated by multiplying the inter quartile value by 1.5 and get it subtracted with R1. The upper inner fence is calculated by multiplying the inter quartile value by 1.5 and add the results with R3. If the lower outer fence is calculated by multiplying the inter quartile value by 3 and get it subtracted with R1. The upper outer fence is calculated by multiplying the inter quartile value by 3 and add the results with R3. The SR of data objects that goes beyond the value of the inner fence is a mild outliers. The SR of data objects that falls outside the outer fence is strong outlier. The results of this method able to find whether the dataset contain a strong, mild or no outliers and it up to the research to exclude the outliers from the execution. In this work, the outlier data objects are identified to be excluded from the clustering.

### 3.4. Percentage Split Distribution (PSD)

Percentage Split is a crucial step of the proposed method, where the data objects are assumed that they are well distributed between the range 0 to 100% in which the min (SR) represents 0% and the max (SR) represents 100%. The clustering portions are then defined by dividing the 100% with number of inputted cluster. This step is an iterative process where each of the iterations calculates the lower and upper limits of each cluster  $C_1 \dots C_t$  using percentage split distribution formula denoted in Equation 6.

$$PSD = ((\text{max}(SR) - \text{min}(SR)) \times \text{percentage}) + \text{min}(SR) \quad (6)$$

Where max (SR) and min (SR) denote maximum and minimum values of SR and percentage represents the split value for each cluster.

### 3.5. If-Then association

If-then Association step gets the lower and upper limits of each cluster from the previous percentage split distribution step and formulates an if-then structure to assign the data objects in to its suitable clusters. SR value of a data object is compared with the

boundary limits of each cluster and assigns the data object into the cluster that it falls under. This step is repeated until all data objects are assigned in a cluster.

### 3.6. SR Silhouette Distance Measure

The internal validation of ILRCPSD algorithm is carried out through silhouette distance measure [9]. But in this work, another novel invention is made with the computation of distance in silhouette measure called SR silhouette Distance Measure, from the traditional way of computing the cluster compactness. SR silhouette distance between and within the clusters are computed using SR values of data objects which means the distance between the data objects are computed using the Manhattan distance of SR values. The distance calculation of SR silhouette measure is denoted in the Equation 7.

$$D_i = \forall_{i=1}^n \forall_{j=1}^n \sum |SR_i - SR_j| \dots \quad (7)$$

Where

i and j are data objects

SR<sub>i</sub> and SR<sub>j</sub> are SR values of i and j

D is the sum of Manhattan distance of data object i with other data objects

The silhouette measure of a data object is the fraction of the subtraction of minimum average distance other clusters with the average distance of its own clusters with the maximum of average of other cluster and own cluster as denoted in Equation 8.

$$SR(i) = \frac{y(i)-x(i)}{\max\{x(i),y(i)\}} \dots \quad (8)$$

Where

i is a data object

y(i) is the minimum average distance of i with other clusters

x(i) is the average distance of i with its own cluster

The simplification of distance calculation with SR values makes the computation task simple and also reduces the operations of square and radical root operations of Euclidean distance.

### 3.7. Rand measure

The external validation of ILRCPSD is obtained through Rand Measure [10] which is the fraction of correctly clustered and correctly rejected data objects with the total number of data objects which can be denoted using the Equation 9.

$$Rand\ measure = \frac{NCCI+NCRI}{TNI} \dots \quad (9)$$

Where

NCCI is the Number of Correctly Clustered Instances

NCRI is the Number of Correctly Rejected Instances

TNI is the Total Number of Instances

## 4. Experimentation and Result Discussion

The experimental study is conducted to analyze the Grade Point Average (GPA) and Intelligent Quotient (IQ) of students to understand the area for improvement to overcome their learning barriers and also to analyze the standards of the proposed contribution through compactness and accuracy. The experimental study is carried out using an IQGPA dataset, collected from [staterunch \[11\]](#). The dataset contains the GPA and IQ scores of 109 students with 6 attributes where the five attributes are the independent variables and the last attribute is the dependent variable. Four classes of students are identified in the datasets as follows:

- Low cognitive and low competency (40)
- Low cognitive and high competency (5)
- High cognitive and low competency (24)
- High cognitive and high competency (40)

The objective of the experiment is to identify the four classes of students using the proposed contributions and to find out the best clustering algorithm that is suitable for clustering the competency and cognitive ability of students.

**Table 1:** IQGPA dataset description

S. No	Attribute Name	Description	Range
1	Math_Score	Mathematics Score	0.0-10.00
2	HS_GPA	Higher Secondary Grade Point Average	0.0-10.00
3	Univ_GPA	University Grade Point Average	0.0-10.00
4	Eng_Score	English Score	0.0-10.00
5	LR_Score	Logical and Reasoning Score	0.0-10.00
6	Class	Categorization of Students	0.0-10.00

The pre-requisite of the experimentation is the investigation of the relationship of attributes of IQGPA which are meaningful for clustering data. The linear relationship between the independent and dependent variables assessed by Pearson correlation method is 0.908561196. Hence, the dataset is processed without any filtration of attributes. Consequently, Table 2 represents the results retrieved from the iterative linear transformation of data through multiple linear regression.

**Table 2:** Iterative Linear Transformation of IQGPA Dataset

Regression Values	Iterative Linear Transformation			
	<i>I1</i>	<i>I2</i>	<i>I3</i>	<i>I4</i>
Multiple R	0.90	0.95	0.96	0.96
R Square	0.82	0.91	0.93	0.92
Adjusted R Square	0.81	0.91	0.92	0.91
Standard Error	0.55	0.38	0.34	0.33
Observations	109	109	109	109

Since, the values of Multiple R, R Square, Adjusted R Square and Standard Error have a decrease the fourth iteration in the linear regression transformation, the iteration is stopped. The co-efficient value that is retrieved for the attributes from the third iteration is multiplied with the data instances to derive the single representation of objects (SR). The SRs are then evaluated against the extreme value analysis method to check the existence of outlier data objects and inputted on to the proposed ILRCPSD algorithms.

The IQGPA dataset is executed over K-means and ILRCPSD algorithms for identifying the most suitable algorithms for clustering similar group of students having similar GPA and IQs scores. The results of the experimental clustering algorithms are validated using time, cluster compactness and cluster accuracy. The goodness of the clustering algorithm is assessed with its ability of grouping the similar data objects with maximized cluster accuracy and cluster compactness.

### 4.1. Time analysis

The time comparative analysis of the experimental algorithms in terms of clustering the students cognitive and competency abilities is presented in Table 3 followed by the pictorial representation in Fig.2. From the table, it has been observed that the proposed ILRCPSD has taken minimum time duration for clustering of data objects with 61 milliseconds than the traditional K-means algorithm with 699 milliseconds.

**Table 3:** Time Analysis - IQGPA

Dataset	Time Taken to Build Clusters in (Milliseconds)	
	<i>K-means</i>	<i>ILRCPSD</i>
IQGPA	699	61

This clearly shows that the proposed contribution is fast enough than the traditional clustering algorithms. Moreover, the results elucidates that the single iterative clustering of data objects using PSD clustering has a direct impact on building clusters with minimized execution time.

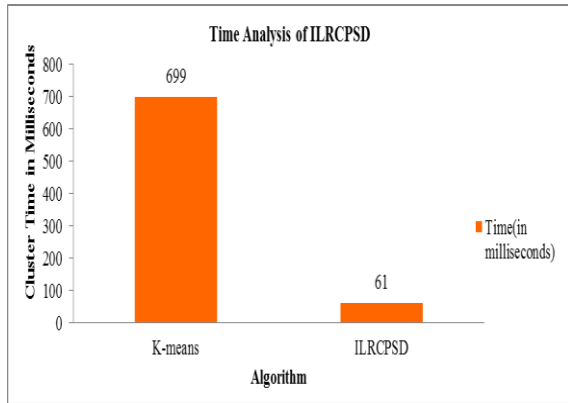


Fig. 2: Time Analysis of ILRCPD

### 4.2. Cluster compactness analysis

The quality of clustering algorithms over the IQGPA dataset is not only assessed by the time factor but also with the compactness of the built clusters. ILRCPD clustering has obtained the highest cluster compactness with 0.616 silhouette similarity between the members of the own clusters and maximum dissimilarity between the members of other clusters. Whereas the cluster compactness achieved by k-means algorithms is lesser than the proposed work. The increase of cluster compactness between the proposed contribution and the traditional k-means is achieved by slicing the single representation of data objects using the percentage split distribution method. Cluster compactness analysis of experimental algorithm is depicted in Table 4 followed by the pictorial representation in Fig. 3.

Table 4: Cluster Compactness Analysis -IQGPA

Dataset	Cluster Compactness (Silhouette Distance)	
	K-means	ILRCPD
IQGPA	0.421	0.616

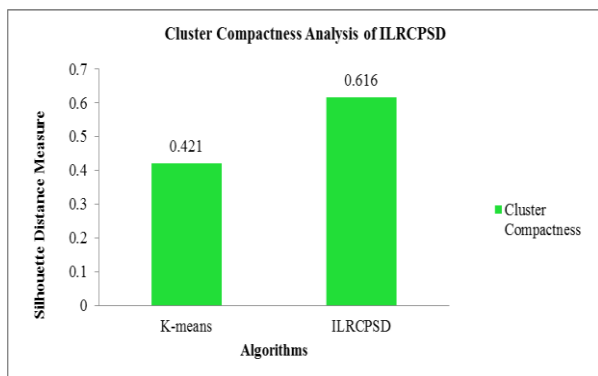


Fig. 3: Cluster Compactness Analysis of ILRCPD

### 4.3. Accuracy analysis

The accuracy of clusters highly depends upon the ability of clustering algorithms in terms of building meaningful clusters. The accuracy analysis is one of the important validation measures for assessing the quality of built clusters. Among the two experimental algorithms shown in Table 5, the ILRCPD algorithm outperforms k-means clustering algorithm with 0.981. Building of meaningful cluster is achievable only when the attributes are mapped with the class label of the dataset. ILRCPD iteratively maps and builds a linear relationship with the labeled classifier which helps the percentage split distribution method slice the clusters to yield the best

cluster accuracy. Fig. 4. denotes the accuracy obtained ILRCPD and K-means.

Table 5: Accuracy Analysis - IQGPA

Dataset	Cluster Compactness (Silhouette Distance)	
	K-means	ILRCPD
IQGPA	0.825	0.981

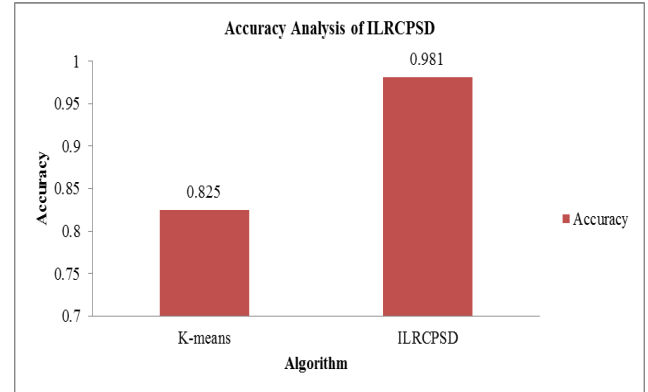


Fig. 4: Accuracy Analysis of ILRCPD

## 5. Conclusion

The real motive of education is not just a successful completion of the academic programme, but it also centres around on how good are the skills of student in terms of placement. The educational institutions are expected to extend their scope in assuring the job career of their students. The assurance of placement depends upon the skills of the students which can be majorly categorized into competency and cognitive skills. Failure in the identification of the background of student's competency and cognitive skills would spoil the career and confidence of the students. Based on the level of skills, the students may be offered more enrichment programmes in order to develop their skills of improvement. Moreover, the implementation of the traditional clustering is challenging in terms of building meaningful clusters. From the experimentation, it has been observed that the traditional K-means clustering algorithm used in the experiment seems to be unsuccessful in grouping the competency and cognitive skills of the higher education students with increased cluster time, low cluster compactness and low cluster accuracy than the other experimental algorithms, because of poor selection of random initial centroids. Thus, to fulfill this research gap, in this study, this research gap is addressed by proposing ILRCPD algorithm. It has been observed that results obtained by the proposed algorithm are remarkable and consistent in terms of time, cluster compactness and accuracy and proved to be best in clustering similar items.

## References

- [1] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40, no. 6 (2010): 601-618.
- [2] Allen, Karen Neuman, and Bruce D. Friedman. "Affective learning: A taxonomy for teaching social work values." Journal of Social Work Values and Ethics 7, no. 2 (2010): 1-12.
- [3] Bolin, A.U., Khramtsova, I. and Saarnio, D., "Using student journals to stimulate authentic learning: Balancing Bloom's cognitive and affective domains." Teaching of Psychology (2005): 154-159.
- [4] Yadav, Jyoti, and Monika Sharma. "A Review of K-mean Algorithm." International Journal of Engineering Trends and Technology (2013): 2972-2976.
- [5] Wu, Xiaohong, Bin Wu, Jun Sun, Shengwei Qiu, and Xiang Li. "A hybrid fuzzy K-harmonic means clustering algorithm." Applied Mathematical Modelling 39 (2015): 3398-3409.

- [6] Prabha, K. Arun, and N. Karthikeyani Visalakshi. "Improved Particle Swarm Optimization Based K-Means Clustering IEEE (2014): 59-63.
- [7] Chawla, Sanjay, and Aristides Gionis. "k-means-: A Unified Approach to Clustering and Outlier Detection." In SIAM, (2013): 189-197.
- [8] Khandare, Anand D. "Modified K-means algorithm for emotional intelligence mining." IEEE (2015): 1-3.
- [9] Kou, Gang, Yi Peng, and Guoxun Wang. "Evaluation of clustering algorithms for financial risk analysis using MCDM methods." Information Sciences 275 (2014):1-12.
- [10] Ansari, Zahid, M. F. Azeem, Waseem Ahmed, and A. Vinaya Babu. "Quantitative evaluation of performance and validity indices for clustering the web navigational sessions." World of Computer Science and Information Technology Journal (2011): 217-226.
- [11] <https://www.statcrunch.com/app/index.php?dataid=1408098>.