ISSN:0975-9646

Gurpreet Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7223-7227

# Proposals Assignment using Fuzzy C-Means and CART Algorithm

Gurpreet Kaur[#1], Jasmeet Singh[*2]

[#]*M.Tech, Research Scholar,*
*Department of Computer Science and Engineering,*
*RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India*

[*]*Assistant Professor,*
*Department of Computer Science and Engineering,*
*RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India*

***Abstract*— The big challenge for many areas such as business, marketing, medical science etc. is management of information. The solution for this is provided by the data mining. The application of data mining is text mining, which provides methods such as classification, clustering etc. to extract the important information from unstructured data or text documents. The technique which is used for grouping similar data objects into one group or cluster is known as clustering. This paper represents proposed method for allotting the proposals according to their quotations like time, technology etc.**

**According to the quotation's decision is taken that how much experienced team is needed for the proposal. The proposals are clustered according to the technology and the experience. This makes easy for work organisation or large companies for assigning proposals according to experience and technology using the data mining techniques. This paper also represents the comparison between proposed and previous work on the basis of different parameters like precision, recall etc.**

***Keywords*— Data mining, Classification, Clustering, Decision tree, CART, Fuzzy c means**

## I. INTRODUCTION

The challenging issue's in today's World is discovering patterns and trends from large databases as the stored information has been increasing day by day. It becomes a major problem for managing and extracting useful information from unstructured data for many areas such as universities, business, research institutes, government funding agencies, and technology intensive companies. The solution for this problem is provided by data mining [1].

The assigning of proposals becomes difficult task for many institutes, organisations, government or private agencies etc. This task may be easy for the small sized organisations [2]. The task becomes difficult for large sized organisations like MNC. The projects which are very important for clients are made by organisations. Thus the experienced persons should prepare the projects. With respect to time and cost, most of projects are lengthy. With respect to the technology or language, the projects are also assigned in which they have to develop. So, the proposals are assigned to well experience team according to the proposal quotations.

**Data mining** emerged in 1980 for creating the useful information. Data mining is used to extract important information from patterns and unknown trends from the large databases. The main goal of the process of data mining is to extract information from a data set and convert this into structure i.e. is easily understandable for further use. Data mining has number of techniques such as classification, clustering, neural networks, decision trees etc. In Data mining i analysis of data is done and with the help of software techniques we finds the patterns and regularities in the set of data mining set [16].

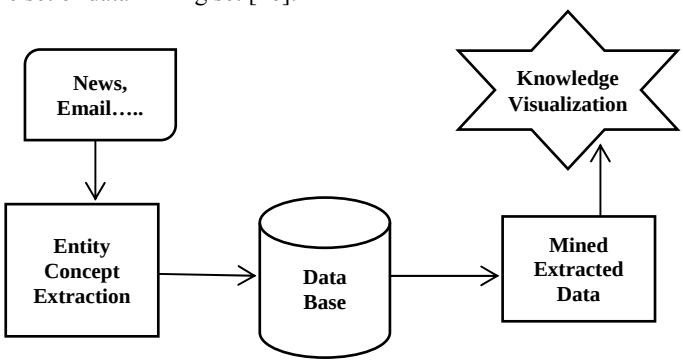


Fig. 1 Data Mining Process

Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The methods used are at the juncture of artificial intelligence, statistics, machine learning, database systems and business intelligence [7].

**Text mining** is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents a step forward from text retrieval. It is a new and different research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. Text mining, sometimes alternately referred to as text data mining, in which we derived the high quality information from text. The main stages of text mining are:

*1) Document pre-processing*: The pre-processing of the document is done to represent the documents in such a way that their storage in the system and retrieval from the system become very efficient. To reduce the length or dimension of the document two methods are used:

 *Filtering*: It is a process for the removal of the words which do not provide any useful or relevant information. Stop word filtering is a standard filtering method.

 *Stemming*: stemming is the process for reducing inflected or derived words to their stem, base or root form. A stemming algorithm reduces the words fishing, fished, fisher to the root word fish.

*2) Text mining technique is applied:* In this stage the text mining algorithm such as classification, clustering, summarization, natural language processing, information extraction is applied.

*3) Text Analysis:* In this stage the outputs which are obtained from the previous steps are analyzed using various tools such as link discovery tool such that user gets important information to achieve the perspective.

**Clustering** provides an important role in a business environment which is an unsupervised learning technique. The grouping of similar types of objects into one cluster is known as clustering. In a same cluster data objects similar to one another remains in same cluster and which are dissimilar remains in other objects [13]. The process of grouping the objects is called cluster analysis in which the can be an abstract such as behaviour of a student, hobbies, handwriting or objects can be physical like a customer [4]. The process which generates a group of objects is known as cluster/s. This cluster consists of the objects which are dissimilar in one cluster and which are similar in one cluster. To find a structure in dataset is the objective of clustering which is exploratory in a nature.

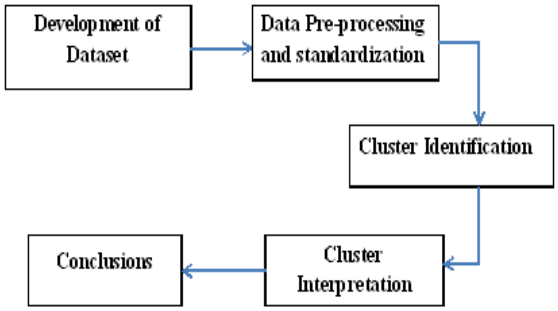


Fig. 2 Clustering Process

After developing the data structure data pre-processing is done. After pre-processing, we identify the clusters and then t clusters are interpreted. At last the conclusion is taken from the process [14]. In data mining, the classifier which is provided by the training data are used to build the classification rules which is a supervised learning technique is known as classification. For unknown classes, test data will predict the values [3]. For a given dataset, there is a problem to select the best classification algorithm. On the basis of their differences, the classification process groups the data into the classes. Most of the classifiers or classification techniques are the Neural Network Classifier, Naïve Bayes Classifier and Decision Tree Classifier and so on. The model that best fits the relationship between the predictors (attributes for prediction) and the prediction (class) use the learning algorithm techniques. The main aim of these techniques is to provide a model which accurately predicts the class of the tuples records or unknown tuples records [5].

**Decision Tree algorithm** is well known for the classification and is very useful. The main advantage is that the process of creating and displaying the results It has an advantage of easy to understand the process of creating and displaying the results [6]. Decision tree can be treated as a tree which means it is a predictive model where branch of the tree is a classification question and leaves represent the partition of the data set according to their classification. Decision Tree is a tree-shaped diagram which shows a statistical probability and is used to determine a course of action [7]. Decision trees used business perspective as original data set is divided through segmentation. There given a data set of attributes with its classes, a decision tree generates a sequence of rules that can be used to recognize the classes for decision making.

## II. Overview of Algorithms

*Fuzzy C-mean Algorithm*

This algorithm is same as k-means algorithm because in this the value of C (number of cluster) has to be defined by the user. Fuzzy C-mean is a technique in which clustering is done by grouping datasets into n clusters. In this every data point belongs to every cluster with a high degree of belonging (connection) to that cluster and other which have low degree of belonging to that cluster lies far away from the centre of a cluster [8]. It is an approach, where the data points have their membership values with the cluster centres, which will be updated iteratively [15]. The FCM algorithm consists of the following steps:

Step 1: Let us suppose that M-dimensional N data points represented by $x_i$ (i = 1, 2, . . . ,N), are to be clustered.

Step 2: Assume the number of clusters that are to be made, that is, C, where $2 \le C \le N$.

Step 3: Select an appropriate level of cluster fuzziness f > 1.

Step 4: Initialize the N × C ×M sized membership matrix U, at random, such that $U_{ijm} \in [0, 1]$ and $P_{Cj=1} U_{ijm} = 1.0$, for each i and a fixed value of m.

Step 5: Determine the cluster centers $CC_{jm}$, for $j_{th}$ cluster and its $m_{th}$ dimension by using the expression given below:

$$CC_{jm} = \frac{\sum^{N}_{i=1} U^{f}_{ijm}\, x_{im}}{\sum^{N}_{i=1} U^{f}_{ijm}}$$

Step 6: Calculate the Euclidean distance between $i_{th}$ data point and $j_{th}$ cluster center with respect to, say $m_{th}$ dimension like the following:

$$D_{ijm} = \left\| (x_{im} - CC_{jm}) \right\|$$

Step 7: Update fuzzy membership matrix U according to Dijm. If Dijm > 0, then

$$Uijm = \frac{1}{\sum^{C}_{c=1} \left( \frac{D_{ijm}}{D_{icm}} \right)^{\frac{2}{f-1}}}$$

If Dijm= 0, then the data point coincides with the corresponding data point of jth cluster center CCjm and it has the full membership value, that is, Uijm = 1.0.
Step 8: Step 5 to Step 7 repeat until the changes in U ≤ ϱ, where ϱ is a pre-specified termination criterion.

*Cart Algorithm*

The CART addresses the classification and regression tree. In this, a binary decision tree is constructed on the basis of some splitting rule i.e. based on the predictor variables. The space of predictor variables is partitioned recursively according to the binary fashion. When a node is reached where no further splitting is possible, repetition of partitioning is stopped. A tree T has a root node whose descendant nodes, called children, further divided into two nodes i.e. terminal nodes and split nodes. The decision trees are built by collection of rules in the modeling data set on the basis of variables as [9]:

- Rules are selected on the basis of variable values to get the best split for differentiate observations on the basis of dependent variables.
- After selecting the rule the node is splits into two branches, and the same process is applied to each branching node.
- Stop splitting when Cart detects that there can be no further gain made.

Each branch of the tree ends into a terminal node. Every terminal node contains one observation and set of rules defined the every terminal node uniquely.

## III. Proposed Work

In previous work, the assignment of proposal to the reviewer is done but not on the basis of proposal quotation and experience. The assignment of the proposals in the developing companies is more important. We worked on appropriate assignment of the proposal using clustering and decision tree algorithms. In this research, the assignment of proposal is done according to the proposal quotations like time period, budget etc. and also the proposals are allotted according to the experience of employees. Also analyzed the results on the basis of various parameters with previous.

The proposed research methodology is divided into following steps:

***Step 1:*** In the first step, we upload proposals that are assigned to developers having experience accordingly.

***Step 2:*** This step will include the implementation of CART algorithm for the classification of the proposals according to the technology. Then on the basis of proposal quotations, the proposals are allotted to members according to their experience.

***Step 3:*** In this step, implementation of Fuzzy c mean algorithm is done for clustering the proposals and employees on the basis of experience and technology.

***Step 4:*** In this step we view the different proposals that are allotted accordingly to the proposal quotations.

***Step 5:*** In the last step, Comparing of the results with the previous results done on the basis of parameters like precision, recall and F-measure and analyze the parameter like execution time, accuracy.

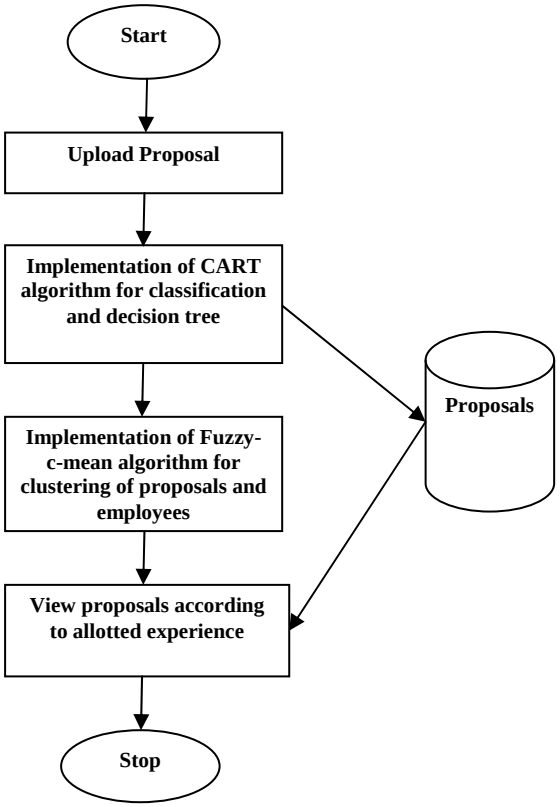


Fig. 3 Flow Diagram

## IV. Results And Comparison

To determine the quality of proposed work, it was necessary to compare it with another algorithm. The performance of the proposed technique is based on the parameters i.e. Precision, Recall, F-Measure, Execution Time.


Fig. 4 Precision Value

TABLE I
COMPARISON OF PRECISION VALUES ON THE BASIS OF NO. OF PROPOSAL'S

| No. of Proposals | Precision of Proposed Work | Precision of Previous Work |
|---|---|---|
| 5 | 0.74 | 0.61 |
| 10 | 0.79 | 0.80 |
| 15 | 0.85 | 0.46 |
| 20 | 0.86 | 0.47 |
| 25 | 0.69 | 0.48 |

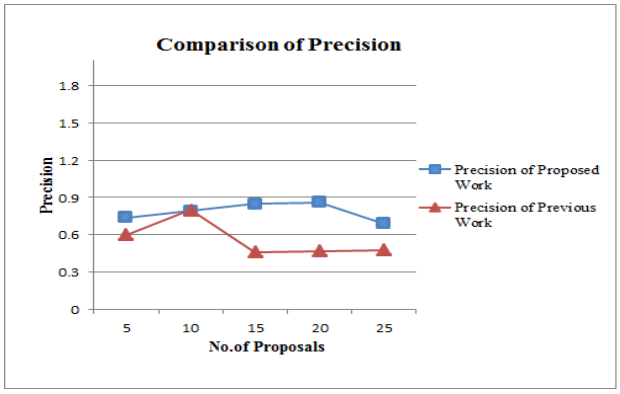


Fig. 5 Comparison of Precision Values

The above graph shows the comparison between precision values of proposed work with the previous work. The precision values of proposed work are better than previous.


Fig. 6 Recall Value

TABLE II
COMPARISON OF RECALL VALUES ON THE BASIS OF NO. OF PROPOSAL'S

| No. of Proposals | Recall of Proposed Work | Recall of Previous Work |
|---|---|---|
| 5 | 0.34 | 0.36 |
| 10 | 0.39 | 0.55 |
| 15 | 0.36 | 0.42 |
| 20 | 0.16 | 0.22 |
| 25 | 0.12 | 0.23 |

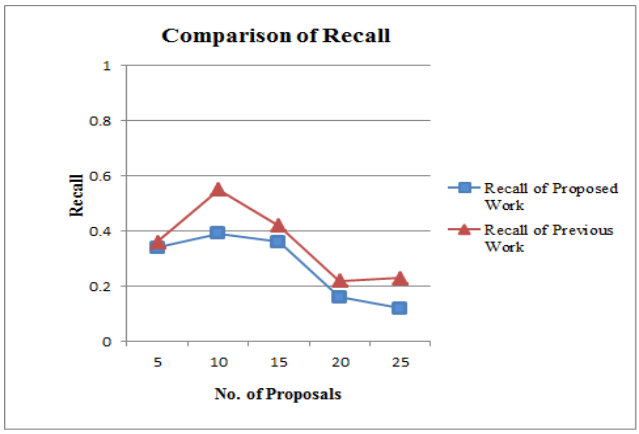


Fig. 7 Comparison of Recall Values

The above graph shows the comparison between recall values of proposed work with the previous work. The precision values of proposed work are better than previous.

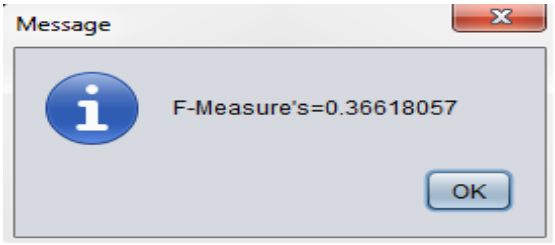


Fig. 8 F-Measure Value

TABLE III
COMPARISON OF F-MEASURE VALUES ON THE BASIS OF NO. OF PROPOSAL'S

| No. of Proposals | F-measure of Proposed Work | F-measure of Previous Work |
|---|---|---|
| 5 | 0.57 | 0.36 |
| 10 | 0.59 | 0.47 |
| 15 | 0.45 | 0.42 |
| 20 | 0.48 | 0.22 |
| 25 | 0.36 | 0.23 |

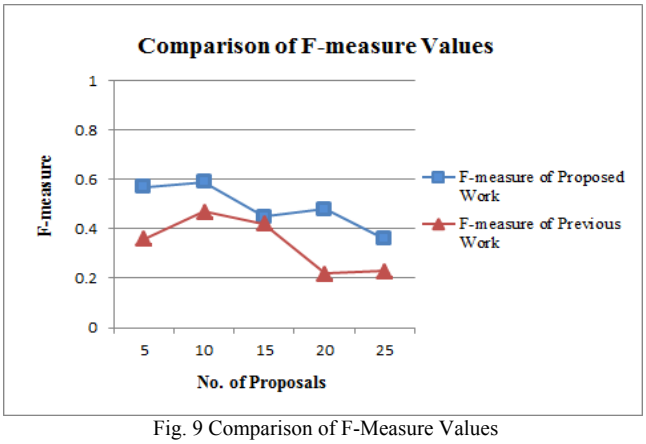


Fig. 9 Comparison of F-Measure Values

The above graph shows the comparison between F-measure values of proposed work with the previous work. The precision values of proposed work are better than previous.

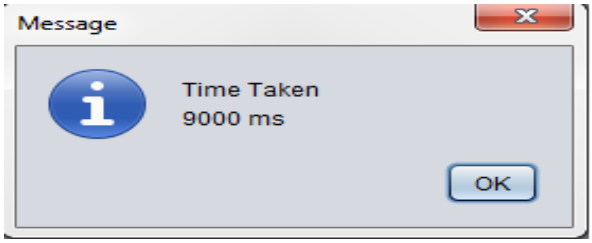


Fig. 10 Execution Time

TABLE IV
EXECUTION TIME ON THE BASIS OF NO. OF PROPOSAL'S

| No. of Proposals | Execution Time (in sec.) |
|---|---|
| 5 | 2 |
| 10 | 4 |
| 15 | 6 |
| 20 | 9 |
| 25 | 11 |

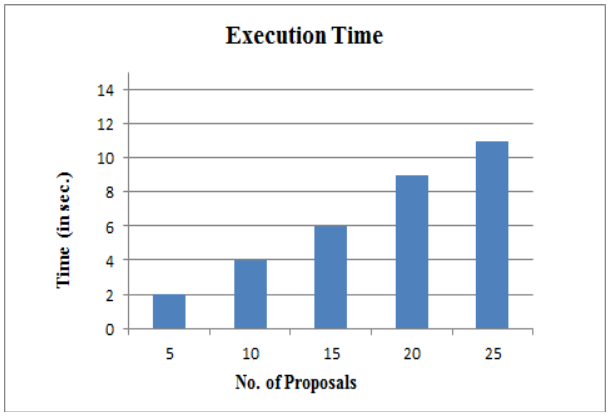


Fig. 11 Representation of Execution Time

The above graph represents the execution time of proposed work.

## V. Conclusion And Future Work

In this paper, we represent a method for assigning the proposal's, according to the experience and technology on the basis of proposal's quotations. The proposed method provides an efficient way for the large organization to assign the proposals using experience factor. In last we compared the proposed method's results with previous using different parameters like precision, recall, F-measure etc.

In future we intended that assigning the proposals to the well-defined teams and also the size of team will be considered.

## Acknowledgment

The author would like to thank the RIMT Institutes, Mandi Gobindgarh-147301, Fatehgarh Sahib, Punjab, India. Author also extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

## References

[1] Divya Nasa, ―*Text Mining Techniques- A Survey*‖, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 2, Issue 4, April 2012, ISSN: 2277 128X pp. 50-54.

[2] Preet Kaur and Richa Sapra ―*Ontology Based Classification and Clustering of Research Proposal and External Research Reviewers*‖, International Journal of Computers & Technology, Volume 5, No. 1, May -June, 2013, ISSN 2277-3061.

[3] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "*Extracting Useful Rules Through Improved Decision Tree Induction Using Information Entropy*", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013.

[4] Prof. Neha Soni, and Prof. Amit Ganatra, "*Categorization of Several Clustering Algorithms from Different Perspective: A Review*", IJARCSSE: International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), vol. 2, Issue 8, August 2012.

[5] Prof. Saurabh Tandel, Prof. Vimal Vaghela, Dr. Nilesh Modi , Dr. Kalpesh Vandra , "*Multi Relational Data Mining Classification Processions – A Survey"*, Int.J.Comp.Tech.Appl,Vol 2 (6), 3097-4002, ISSN:2229-6093.

[6] T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.Prasanna Venkatesan , "*An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data*", I.J.Modern Education and Computer Science, 2013, 5, 18-27 Published Online June 2013 in MECS.

[7] Madhuri V. Joseph, Lipsa Sadath, Vanaja Rajan, "*Data Mining: A Comparative Study on Various Techniques and Methods*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 3, Issue 2, February 2013, pp. 106-113.

[8] http://en.m.wikipedia.org/wiki/fuzzy_clustering.

[9] David G.T. Dension, Bani K. Mallick, Adrian F.M. Smith, "*A Bayesian CART Algorithm*", Biometrika, Vol. 85, No. 2 (Jun., 1998), 363-377.

[10] N.Arunachalam, E.Sathya ,S.Hismath Begum and M.Uma Makeswari "An Ontology Based Text Mining Framework for R&D Project Selection" published in International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 1, February 2013.

[11] Jay Prakash Pandey, Shrikant Lade, Manish Kumar Suman "Automatic Soumi Ghosh and Sanjay Kumar Dubey , **"**Comparative Analysis of K-means and Fuzzy C-Means Algorithms**"**, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[12] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, ―*Enhanced K-Mean Clustering Algorithm To Reduce Number Of Iterations And Time Complexity*‖, Middle-East Journal of Scientific Research 12 (7): 959-963, 2012, ISSN 1990-9233.

[13] Shaidah Jusoh and Hejab M. Alfawareh**, "**Techniques, Applications and Challenging Issue in Text Mining**",** IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012, ISSN (Online): 1694-0814.

[14] Ontology Creation for Research paper classification" published in International Journal of Engineering Research and Science & Technology, Vol. 2, No. 4, November 2013.

[15] Hmway Hmway Tar, Thi Thi Soe Nyunt "Ontology-Based Concept Weighting for Text Documents", International Conference on Information Communication and Management IPCSIT vol.16 2011 IACSIT Press, Singapore.

[16] Amandeep Kaur Mann, and Navneet Kaur, "Survey Paper on Clustering Techniques", IJSETR: International Journal of Science, Engineering and Technology Research (ISSN: 2278-7798), vol. 2, Issue 4, April 2013.

[17] http://www.intechopen.com/books/applications-of-self-organising-maps.