IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 18-20

Introduction On Selecting A Pair From A Group Of Objects Using Data Mining: A Literature Survey

Er. Sampada Pongade[1], Prof. Vanita Tonge[2]
[1]Er. Sampada Pongade, Department of CSE, RCERT , Chandrapur, Maharashtra, India
[2] Prof. Vanita Tonge, Department of CSE, RCERT, Chandrapur, Maharashtra, India

***Abstract**: The paper is based on research area data mining in computer science. Data mining means knowledge mining from data. From among different approaches in data mining we are going to work on a approach which is "Proximity Measures for Binary Attributes". The paper is about selecting a single pair from all the possible pairs in the data set consisting of objects of same type and having some attributes. Based on those attributes which will be binary in nature the research component i.e. Distance Measure will be calculated and a single pair will be given as output.*

I. Introduction

Data mining sometimes called data or knowledge discovery in databases is the extraction of hidden predictive information from large databases [1]. Data mining field uses many methods to extract the needed hidden data and hidden patterns from big data[2]. Text data mining resembles data mining because it extracts useful knowledge and information by analyzing the diversified viewpoints of written data [3]. The term data mining was originally used to describe the process through which previously unknown patterns in data were discovered[4]. Objects in this paper are the names of officers recruited by UPSC and serving in India . The two officers should be either IAS/IPS/IRS and also their status should be married to get a single cadre allotted. The single cadre allotment policy is specified by our honourable PM Shri. Narendra Modiji(till 2019). According to him, two IAS/IPS/IRS officers who are married can be allotted a single cadre i.e. state for serving in India until their retirement period unless suspended or resigned that too on certain grounds. Names of officers recruited by UPSC will be consisting of UPSC Results of batch 2004-2016. List of these officers will be shortlisted by the criteria that they should be either IAS/IPS/IRS. This shortlisted list will be finally used for selecting a pair, so that a single cadre could be allotted to that pair which we will get as output on applying the "Proximity measure for binary attributes approach" in data mining.

Proximity measure for binary attributes approach:
A contingency table for binary data is created as follows:

	Object j			
Object i		1	0	Sum
	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p=q+r+s+t

Distance measure for binary attributes:
$d(i,j) = (r+s)/(q+r+s)$
Objects having more distance measure value are more similar.

II. Literature Survey

"Is Alcohol Affect Higher Education Students Performance: Searching and Predicting pattern using Data Mining Algorithms"

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node(non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node(or terminal node) holds a class label. The topmost node in a tree is the root node. It represents the concept buys_computer, that is, it predicts whether a customer at AllElectronics is likely to purchase a computer. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees(where each internal node branches to exactly two other nodes), whereas others can produce nonbinary trees.

In this experimental dataset used data set about MCA student on their courses which holds 450 instances. Four Decisions Tree algorithms (BFTree, J48, RepTree and Simple Cart) are applied in this work.

The results showed that BFTree algorithm mostly proper to classify and predict student's whose performance is excellent and who's poor during studying the subjects. [1]

"Analysis of Mahatma Gandhi National Rural Employment Guarantee Act Using Data Mining Technique"

Data visualization aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications-for example, at work for reporting, managing business operations, and tracking progress of tasks. More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data. Nowadays, people also use data visualization to create fun and interesting graphics.

The Mahatma Gandhi NREGA sponsors various schemes for helping rural people below the poverty-line for creation of wage employment and productive assets. This paper gives the analysis of the payment of wages to the workers under MGNREG scheme in districts of Rajasthan, using decision tree J48 classification technique. [2]

"Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph"

Methods for mining the simplest form of frequent patterns are such as those discussed for market basket analysis. Apriori is the basic algorithm for finding frequent itemsets. We can generate strong association rules from frequent itemsets. Several variations to the Apriori algorithm for improved efficiency and scalability. We can also apply pattern-growth methods for mining frequent itemsets that confine the subsequent search space to only the data sets containing the current frequent itemsets.

In the present study, to classify the vast amount of Care Life Log data that occurs in nursing in one Miyazaki Hospital Long-term Health Care Facility by level of care required, data mining was carried out. The characteristic vocabulary from the Long-term Health Care Facility's Care Life Log was used to integrate and analyze the level of care required. [3]

"Introduction to Data, Text, and Web Mining for Business Analytics Minitrack"

Support vector machines(SVMs), a method for the classification of both linear and nonlinear data. In a nutshell, an SVM is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane(i.e., a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors("essential" training tuples) and margins(defined by the support vectors).

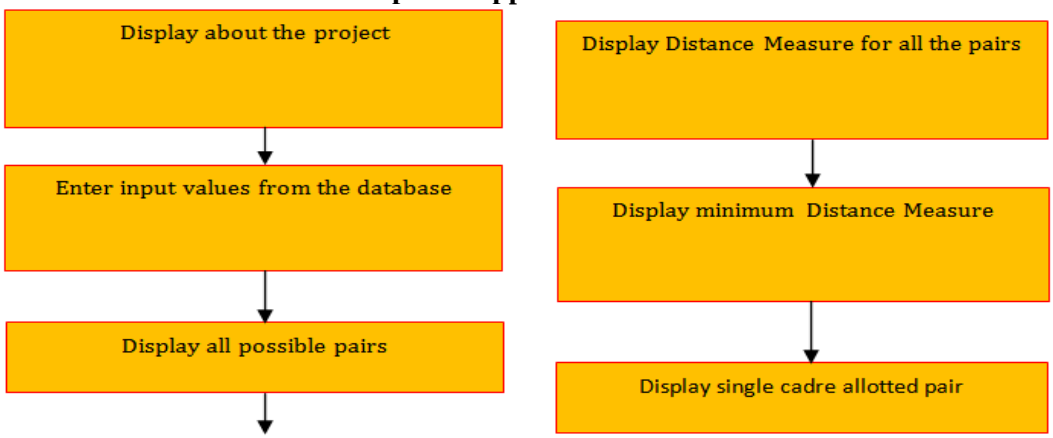
This method is used for extracting and providing much needed insight and knowledge to decision makers at any and all levels of managerial hierarchy. [4]

"Effieient Algorithms to find Frequent Itemset Using Data Mining"

Differentially private data mining algorithms shows more interest because data item mining is most facing problem in data mining. It is useful in most applications like decision support, Web usage mining, bioinformatics, etc.

To achieve privacy, utility and efficiency Frequent Itemset Mining algorithm is proposed which is based on the Frequent Pattern growth algorithm. Private Frequent Pattern -growth algorithm is divided into two phases namely preprocessing phase and Mining phase. [5]

III. Proposed Approach Workflow :



This module will contain the details of what the project is about in brief.

Module 2-Enter input values from the database :

This module will enable a user to enter the input values from the database created and display those input values.

Module 3-Display all possible pairs :

This module contains all the objects to form all possible pairs and display those pairs.

Module 4-Display Distance Measure for all the pairs :

This module will display value of Distance Measure for all the pairs.

Module 5-Display minimum Distance Measure

This module will display minimum value of Distance Measure out of all the Distance Measure values.

Module 6-Display single cadre allotted pair

This module will display final output i.e. single cadre allotted pair.

IV. Conclusion

I have studied the "Proximity Measures for Binary Attributes" in detail. The research component for this project is "Distance Measure". Using the research component we will get the final output as the single cadre allotted pair.

I am going to apply the above mentioned approach to allot a single cadre for a pair out of all the possible pairs in the data set of objects i.e. the names of UPSC Results of batch 2004-2016.

References

[1]. Saurabh Pal, "Is Alcohol Affect Higher Education Students Performance : Searching and Predicting pattern using Data Mining Algorithms", International Research Journal of Engineering and Technology (IRJET)

[2]. Kritika Yadav, "Analysis of Mahatma Gandhi National Rural Employment Guarantee Act using Data Mining Technique", International Journal of Computational Intelligence Research (IJCIR)

[3]. Muneo Kushima, Kenji Araki, Tomoyoshi Yamazaki, Sanae Araki, Taisuke Ogawa, Noboru Sonehara, "Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph", Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I, (IMECS) 2017

[4]. Dursun Delen, Enes Eryarsoy, Şadi E. Şeker, "Introduction to Data, Text, and Web Mining for Business Analytics Minitrack", Proceedings of the 50th Hawaii International Conference on System Sciences | 2017

[5]. Sagar Bhise, "Effieient Algorithms to find Frequent Itemset Using Data Mining", International Research Journal of Engineering and Technology (IRJET)

BIOGRAPHIES

Er. Sampada Pongade
B.E. (Computer Technology)
Mtech pursuing (Computer Science and Engineering)