Implementation of Proximity measure approach for binary attributes to select a pair from a group of objects

[1]Er. Sampada Pongade, [2]Prof. Vanita Tonge

CSE Department
RCERT
Chandrapur, 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.**

Keywords*: *Data mining; Proximity measure approach for binary attributes; Distance Measure; objects; attributes

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]. The term data mining was originally used to describe the process through which previously unknown patterns in data were discovered. This definition has since been stretched beyond those limits by software vendors and consultancy companies to include most forms of data analysis in order to increase its reach and capability. With the emergence of analytics as an overarching term for all data analyses, data mining is put back into its proper place—a critical part of analytics continuum where the new discovery of knowledge happens.[5]

Dataset consisting of names of civil servants will be used as input , so that a single cadre could be allotted to that pair which we will get as output on applying the "Proximity measure approach for binary attributes" in data mining but the approach is applied after processing the dataset according to "Single cadre allotment policy" which is proposed by Hon. PM Shri Narendra Modi.

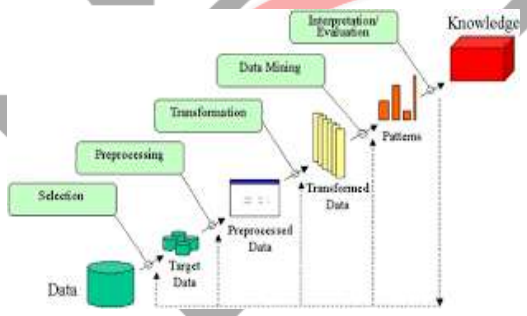


Fig. Knowledge Discovery Process (Data mining)

The knowledge discovery process is shown in above Fig. as an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

In data mining applications, such as clustering, outlier analysis, and nearest-neighbor classification, we need ways to assess how alike or unalike objects are in comparison to one another. For example, a store may want to search for clusters of customer objects, resulting in groups of customers with similar characteristics (e.g., similar income, area of residence, and age). Such information can then be used for marketing. A cluster is a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters. Outlier analysis also employs clustering-based techniques to identify potential outliers as objects that are highly dissimilar to others. Knowledge of object similarities can also be used in nearest-neighbor classification schemes where a given object (e.g., a patient) is assigned a class label (relating to, say, a diagnosis) based on its similarity toward other objects in the model.

Similarity and dissimilarity measures, which are referred to as measures of proximity. Similarity and dissimilarity are related. A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unalike. The higher the similarity value, the greater the similarity between objects. (Typically, a value of 1 indicates complete similarity, that is, the objects are identical.) A dissimilarity measure works the opposite way. It returns a value of 0 if the objects are the same (and therefore, far from being dissimilar). The higher the dissimilarity value, the more dissimilar the two objects are.

II. PROBLEM DEFINITION

A. Details of Problem Definition

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. 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.

B. Data Set Used

Civil servants recruited by UPSC consisting of batch 2004-2016 is used as data set. This dataset is taken from a magazine named "YOJANA" which is published by government every month especially for the civil servant aspirants.

This data set is used as input and finally we get an officer's name who has been allotted a single cadre for service until his/her retirement period.

C. Distance measure for different types of data

C1. Proximity Measures for Nominal Attributes

A nominal attribute can take on two or more states. For example, map color is a nominal attribute that may have, say, five states: red, yellow, green, pink and blue.

Let the number of states of a nominal attribute be M. The states can be denoted by letters, symbols, or a set of integers, such as 1, 2,…, M. Notice that such integers are used just for data handling and do not represent any specific ordering.

"How is dissimilarity computed between objects described by nominal attributes?"

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$d(i,j)=(p-m)/p$, where m is the number of matches (i.e., the number of attributes for which i and j are in the same state), and p is the total number of attributes describing the objects. Weights can be assigned to increase the effect of m or to assign greater weight to the matches in attributes having a larger number of states.

C2. Proximity Measures for Binary Attributes

Let's look at dissimilarity and similarity measures for objects described by either symmetric or asymmetric binary attributes.

A binary attribute has only one of two states: 0 and 1, where 0 means that the attribute is absent, and 1 means that it is present. Given the attribute *smoker* describing a patient, for instance, 1 indicates that the patient smokes, while 0 indicates that the patient does not. Treating binary attributes as if they are numeric can be misleading. Therefore, methods specific to binary data are necessary for computing dissimilarity.

"So, how can we compute the dissimilarity between two binary attributes?"

One approach involves computing a dissimilarity matrix from the given binary data. If all binary attributes are thought of as having the same weight, we have the 2*2 contingency table of Table 1, where *q* is the number of attributes that equal 1 for both objects i and j, *r* is the number of attributes that equal 1 for object i but equal 0 for object j, s is the number of attributes that equal 0 for object i but equal 1 for object j, and *t* is the number of attributes that equal 0 for both objects *i* and *j*. The total number of attributes is *p*, where $p = q+r+s+t$.

For symmetric binary attributes, each state is equally valuable. Dissimilarity that is based on symmetric binary attributes is called symmetric binary dissimilarity. If objects i and j are described by symmetric binary attributes, then the

Table 1. Contingency Table for Binary Attributes

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

dissimilarity between i and j is d(i,j)=(r+s)/(q+r+s+t)

For asymmetric binary attributes, the two states are not equally important, such as the positive (1) and negative (0) outcomes of a disease test. Given two asymmetric binary attributes, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match). Therefore, such binary attributes are often considered "monary" (having one state). The dissimilarity based on these attributes is called asymmetric binary dissimilarity, where the number of negative matches, t , is considered unimportant and is thus ignored in the following computation:

d(i,j)=(r+s)/(q+r+s)

C3. Dissimilarity of Numeric Data: Minkowski Distance

Here, we describe distance measures that are commonly used for computing the dissimilarity of objects described by numeric attributes. These measures include the Euclidean, Manhattan, and Minkowski distances.

In some cases, the data are normalized before applying distance calculations. This involves transforming the data to fall within a smaller or common range, such as [-1, 1] or [0.0, 1.0]. Consider a height attribute, for example, which could be measured in either meters or inches. In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such attributes greater effect or "weight."

Normalizing the data attempts to give all attributes an equal weight. It may or may not be useful in a particular application.

The most popular distance measure is Euclidean distance (i.e., straight line or "as the crow flies"). Let i =(.xi1, xi2, … , xip) and j =(xj1, xj2, … , xjp) be two objects described by *p* numeric attributes. The Euclidean distance between objects *i* and *j* is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

Another well-known measure is the Manhattan (or city block) distance, named so

because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

D. Methodology Used

Proximity Measure Approach for asymmetric binary attributes :

A contingency table for binary data is created as follows:

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

Fig. Contingency Table

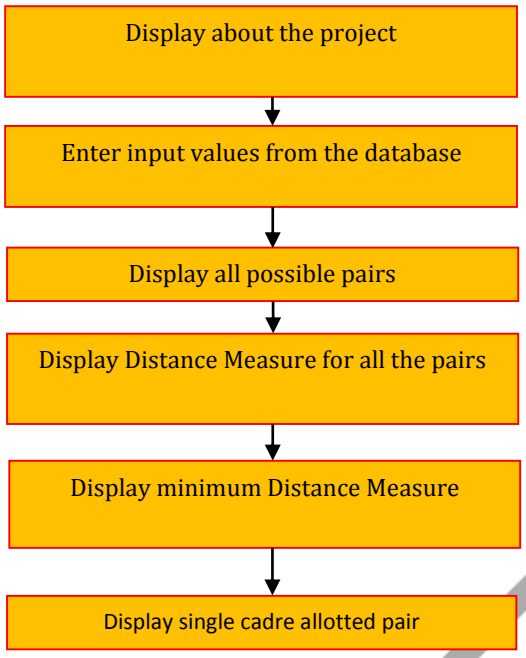
Distance measure for asymmetric binary attributes:

d(i,j) = (r+s)/(q+r+s)

Objects having more distance measure value are more similar.

III. DETAILS OF IMPLEMENTATION

A. Flowchart of overall work



B. Proposed Algorithm

1. Start
2. Show about the project details in short
3. Show the input values i.e. names of IAS/IFS/IRS officers from the MS-Access database
4. Show the eligible officers names
5. Calculate the Distance Measure value for all the officers pairs and show it one by one
6. Compare all the Distance Measure values

a) If all the values of Distance Measure are same then

i) Show the Distance Measure value

ii) Compare the AIR value and find the min. AIR value

Otherwise

i) Show the minimum Distance Measure value

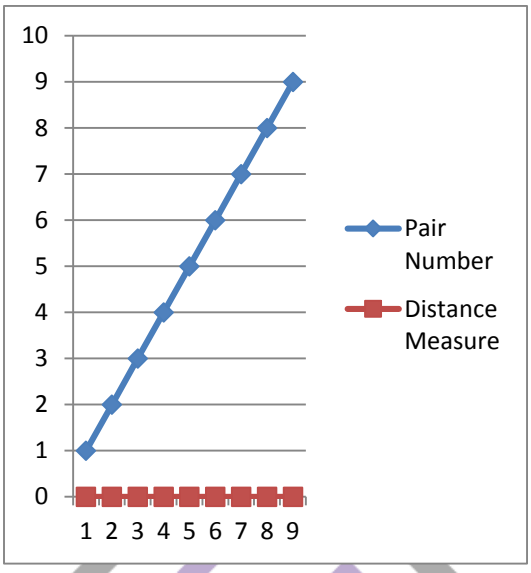
7. If all the values of Distance Measure are same then show the name of officer whose pair is allotted a single cadre based on min. AIR value
8. Stop

C. Result Analysis

The main Research component is Distance Measure. In this research work all the Distance Measure values came out to be 0.So it required to make one more step further i.e. to check AIR (All India Rank) of the eligible civil servants. Min. value of AIR is considered to be highest merit rank. So after comparing the AIR values of all the eligible candidates "Gaurav Agrawal" with AIR 1 is found as the final result of the research work.

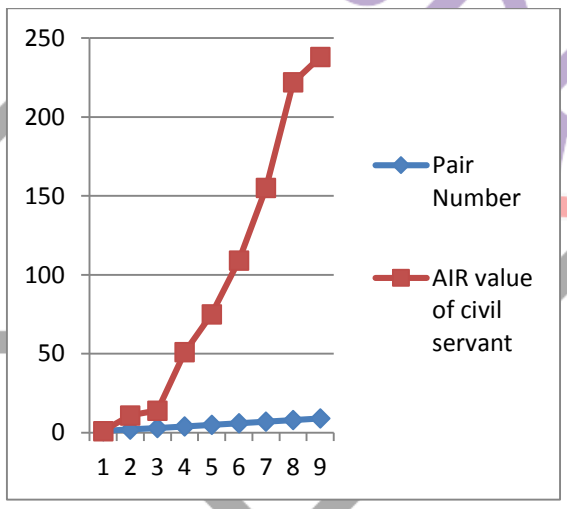
IV. DATA ANALYSIS

The main research component is Distance Measure. But the value of Distance Measure is same i.e. 0 for all the pairs.



So, now there is a need to check for another attribute. That attribute value is AIR (All India Rank). Now, we check for AIR of all the officers corresponding to all the 9 pairs.

Now, whatever may be the AIR value, we need to focus on minimum AIR as it shows most capable officer i.e. on the basis of merit.



From the above graph, it can be observed that minimum AIR value is for Pair1.

Now, the corresponding officer's name for Pair1 is Gaurav Agrawal which is the final output of the overall research work.

V. CONCLUSION

I have executed the project successfully and thus got "Gaurav Agrawal" with AIR 1 who can be allotted a single cadre for service in India uptil his retirement as output. This research work is done using "Proximity measure approach for binary attributes" in data mining research area.

ACKNOWLEDGMENT

Thanks to the government of India who publishes the **"Yojana" magazine** from where I selected the dataset.

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)

[6] Vishal Bhemwala, Bhavesh Patel, Dr. Ashok Patel, "Distributed Data Mining: Implementing Data Mining Jobs on Grid Environments", 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099

[7] Charles C.N. Wang, Jeffrey J P Tsai, "Application of Semantic Computing in Cancer on Secondary Data Analysis", Conference Paper · January 2018

[8] Akshat Savaliya, Aakash Bhatia, Jitendra Bhatia, "Application of Data Mining Techniques in IoT: A Short Review", 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099 National Conference on Advanced Research Trends in Information and Computing Technologies (NCARTICT-2018)

[9] Dr.Mohd Ashraf, Dr. Zair Hussain, " Investigation of Performance Analysis of Classification Algorithm in Data Mining", 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099

[10] Sneha Sakharkar, Shubhangi Karnuke, Snehal Doifode, Vaishnavi Deshmukh, "A Research Homomorphic Encryption Scheme to Secure Data Mining in Cloud Computing for Banking System", 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099

[11] Sheng-Yong Niu, Jinyu Yang, Adam McDermaid, Jing Zhao, Yu Kang and Qin Ma, "Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes", Briefings in Bioinformatics, 19(2), 2018, 360

[12] T. Kesava, G. Mohan Ram, " A Novel Sorting Algorithm for Data Analysis", International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

[13] Jignasha M. Jethva, Nikhil Gondaliya, Vinita Shah, " A Review on Data Mining Techniques for Fertilizer Recommendation", International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

[14] L.T. Kell, A. Ben Mhamed, T. Rouyer and A. Kimoto, "AN EXPLORATORY DATA ANALYSIS OF THE EAST ATLANTIC BLUEFIN STOCK ASSESSMENT DATASET", Collect. Vol. Sci. Pap. ICCAT, 74(6): 3037-3051 (2018)

[15] Mikhail Kanevski and Mohamed Laib, "Analysis of high dimensional environmental data using local fractality concept and machine learning", Geophysical Research Abstracts Vol. 20, EGU2018-4405, 2018 EGU General Assembly 2018

[16] Ludovic Duponchel, "When remote sensing meets topological data analysis", Article February 2018 at: https://www.researchgate.net/publication/322959542

[17] Ujwal UJ, Dr Antony PJ and Sachin DN, " Predictive Analysis of Sports Data using Google Prediction API", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2814-2816

[18] Paul Rosen , Mustafa Hajij , Junyi Tu , Tanvirul Arafin , Les Piegl, "Inferring Quality in Point Cloud-based 3D Printed Objects using Topological Data Analysis", Article · January 2018 at: https://www.researchgate.net/publication/322537346

[19] Sartaj Ahmad and Rishabh Varma, "Information extraction from text messages using data mining techniques", Malaya Journal of Matematik, Vol. S, No. 1, 26-29, 2018

[20] Risto Vaarandi and Mauno Pihelgas, "LogCluster – A Data Clustering and Pattern Mining Algorithm for Event Logs", 11th International Conference on Network and Service Management (CNSM 2015), ISBN: 978-3-901882-77-7

[21] A.K.Dubey, Alakh N Sharma, Balwant Singh Mehta, Sunita Sanghi, "Empowering Youth", Yojana magazine-June 2017