VAT for Preference Data

Anukarsh Khandelwal | Yash Agrawal | Prof. Dheeraj Kumar

Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee

ABSTRACT

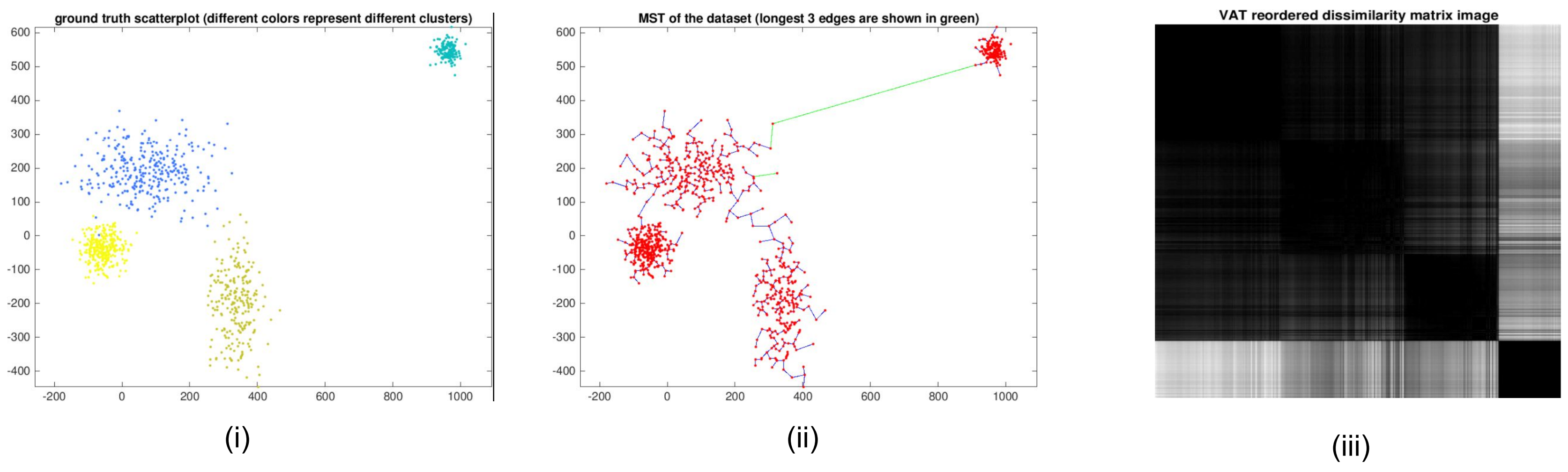
Visual Assessment Tendency(VAT) is a method used for visually assessing the number of clusters in a given dataset. Conventional VAT fails to give correct output on many cases of preference data. In this project, we aim to modify VAT to deal directly with preference data. We use Chu–Liu/Edmond's Algorithm for reordering preference matrices which is an essential step for getting VAT.

INTRODUCTION

- In today's time where data is available in every form and factor, understanding these data help in providing meaningful information for making better decisions. Clustering is an important way of understanding an unlabelled data.
- A method is given for visually assessing the clustering tendency of a set of objects when they are represented either as object vectors or by numerical pairwise dissimilarity values. The objects are reordered and the reordered matrix of pairwise object dissimilarities is displayed as an intensity image. Clusters are indicated by dark blocks of pixels along the diagonal.
- Preference Dataset is a type of dataset where only relative information is given between some/all pair of indexes. A particular cell (P_{ij}) in a preference matrix (P) represents the preference of 'I' over 'J'.

THEORY

(i) is scatterplot for a randomly generated 2-dimensional synthetic dataset having a total of 1,000,000 points, distributed among 4 clusters with ground truth values.
(ii) shows Minimum Spanning Tree generated by VAT calculated using Euclidean distance between data points. The three longest edges, which is used for separating clusters, are shown in green, This MST is used by VAT to reorder dissimilarity matrix for this dataset.
(iii) is the output of VAT on this dataset, is an image matrix where the colour of a cell is a reflection of value in its respective cell. Clustering tendency can be easily assessed from the received image. Broadly, it can be divided into two clusters. Looking closely in the bigger square, three different dark squares can be seen.



(i) (ii) (iii)

Given a set of n options, a fuzzy preference relation is specified by an n x n fuzzy preference matrix where each matrix element p_{ij} quantifies the degree of preference of option i over j.

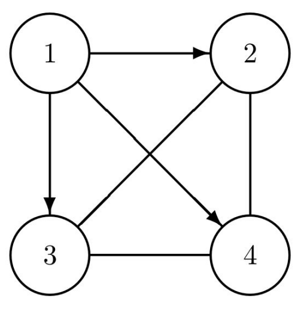
$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}$$

Here we consider additive fuzzy preference matrices with $p_{ij} \in [0, 1]$, where $p_{ij} = 0$ indicates absolutely no preference, $p_{ij} = 1$ indicates complete preference, and $p_{ij} = ½$ indicate equivalence.

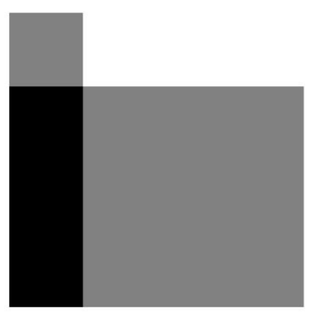
A straightforward way to convert such a reciprocal additive fuzzy preference matrix into a symmetric dissimilarity matrix is by : $\mathbf{d_{ij} = d_{ji} = max(p_{ij} , P_{ji}) - 0.5}$ …(1)

Good Example: Considering a preference matrix P1, we can calculate its dissimilarity matrix D1 using the aforementioned formula. P1 can be translated to preference graph G1.

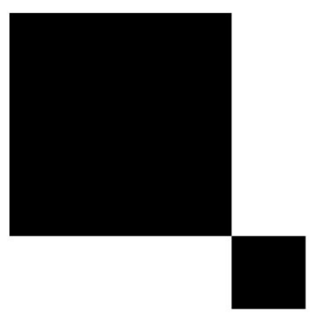
$$P_1 = \begin{pmatrix} 0.5 & 1 & 1 & 1 \\ 0 & 0.5 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 & 0.5 \end{pmatrix} \quad D_1 = \begin{pmatrix} 0 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \end{pmatrix}$$



VAT results for the P1 and D1 comes out to be fig. 1 and fig. 2 respectively. Fig. 1 fails to give any useful information whereas fig. 2 clearly shows presence of two clusters, which is the expected outcome.



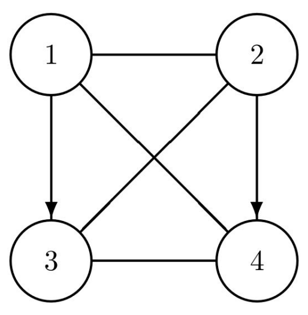
(fig. 1)



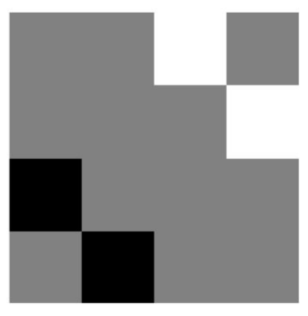
(fig. 2)

Bad Example: Now considering a preference matrix P2 and D2 is made using the formula (i), we get fig. 3 as result from P2 and fig. 4 as result from D2.

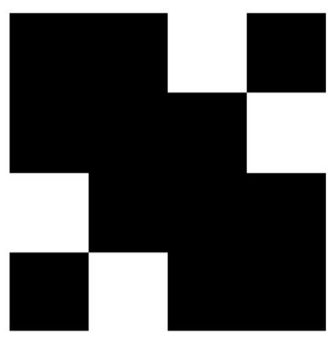
$$P_2 = \begin{pmatrix} 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \\ 0 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 & 0.5 \end{pmatrix} \quad D_2 = \begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \end{pmatrix}$$



Both fig. 1 and fig. 2 do no yield expected results. Many more such cases can be made where VAT does not yeild images that appropriately reflect the preference structure.



(fig. 3)

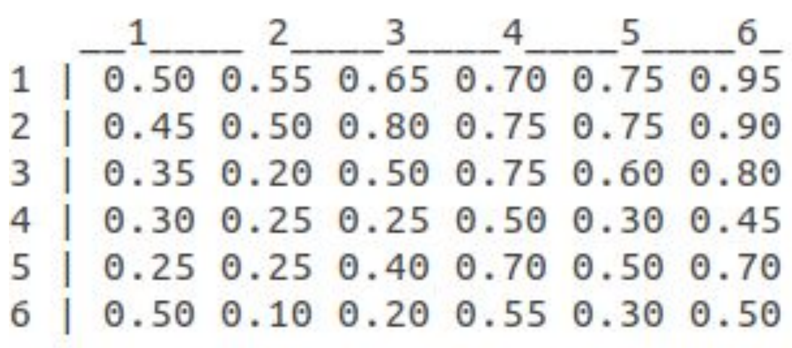


(fig. 4)

METHODS

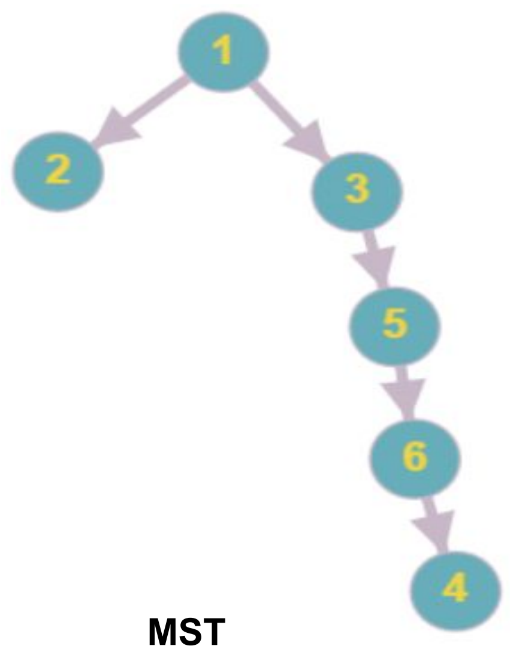
- In conventional VAT, we have values of n features for every node where pairwise euclidean distance between any two nodes can easily be calculated, whereas in preference matrix we don't have any ground truth and only relative information is available.
- This is analogues to conventional VAT with a difference that the edges here are directed. We know that a node is preferred over another by some value. Similar to conventional VAT we can find MST for reordering and VAT can be visualized.
- **Edmonds' algorithm** or **Chu–Liu** is an algorithm for finding a spanning arborescence(for a vertex *u* called the root and any other vertex *v*, there is exactly one directed path from *u* to *v*) of maximum weight. It is the directed analogue of the minimum spanning tree problem.
- The preference matrix is now converted into a graph and we can apply Edmond's algorithm on it considering the fact all other nodes are can be visited form the assumed root. We consider the root as the object which has the maximum priority over all other objects. A node having the maximum number of outgoing preferred edges going out from it is considered as root.
- The weights between any two nodes 'a' to 'b' were replaced by value equal to **x / |x - 0.5|** where x is the preference value of 'a' over 'b', this value always lies between 0 and 1. This change in weight assign edges connecting two equally preferred edges high values, and absolutely preferred edges will have least values.
- We use resultant MST to find pairwise distance between every nodes and a heatmap is plotted according to this value, where 0 is the darkest and 1 is the lightest colour. The resultant image can be visualized for assessment of clustering tendency.

RESULTS

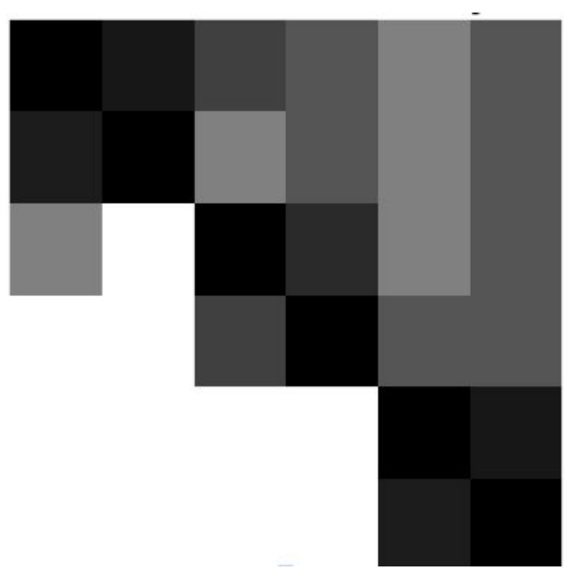


```
  __1____ 2____3____4____5____6_
1 | 0.50 0.55 0.65 0.70 0.75 0.95
2 | 0.45 0.50 0.80 0.75 0.75 0.90
3 | 0.35 0.20 0.50 0.75 0.60 0.80
4 | 0.30 0.25 0.25 0.50 0.30 0.45
5 | 0.25 0.25 0.40 0.70 0.50 0.70
6 | 0.50 0.10 0.20 0.55 0.30 0.50
```

PREFERENCE MATRIX



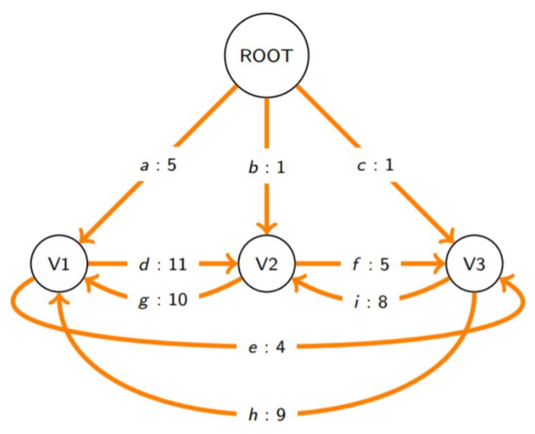
MST



EDMOND'S VAT REORDERED MATRIX IMAGE

We consider a case of 6 teams playing against each other and their probability of a team winning over another is given in the preference matrix. After applying our approach on the preference matrix we got the MST in the order I = [1, 2, 3, 5, 6, 4]. Reordering the weight matrix according to order I and the resultant matrix can then be visualized as VAT image. The obtained result shows that there are there almost three clusters formed (1,2), (3,5) and (6,4) which is expected. Studying the preference matrix 'Team 1' and 'Team 2' have a higher probability of winning against other teams. 'Team 6' and 'Team 4' have a low probability of winning. 'Team 3' and 'Team 5' show average performance.

IMPROVEMENTS



- One of the case where the Edmond's algorithm will not work for our case is shown. Here for the maximum sum path there are two different paths which have an equal sum of weights. 1)Root->V1->V2->V3 2)Root->V3->V1->V2. Path 2 is actually kind of wrong due to less weight between ROOT and V3.
- If from one node we have directed edges having same weights, then the node in which we have proceed first is an issue for directed edges>=3. For two edges we can see if there is a parent-child relationship between them.

REFERENCES

[1] D. Kumar, J. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. Havens, "A Hybrid Approach to Clustering in Big Data." IEEE Transactions on Cybernetics, vol. 46, no. 10, pp. 2372-2385, Oct. 2016.

[2] James C.Bezdek, Bonnie Spillman, Richard Spillman, "A fuzzy relation space for group decision theory."Fuzzy Sets and Systems Volume 1, Issue 4, October 1978, Pages 255-268.