The study of the data context map of fusing attributes into a unified display

***Abstract*—For the visualisation of a data matrix, which is a multidimensional matrix of data, there have been a large number of methods described earlier. However, they all suffer from a very common problem: that the observation of these data points with respect to the attributes is inaccurate. In the given paper, the authors Cheng and Mueller [1] have discussed a method that allows all types of comprehensive layouts. These methods described were a combination of two similarity matrices that were used in isolation, one for similarity of attributes and the other for data points. This combined matrix is required for a full multi-dimensional scaling type layout.The resulting layout, which places the data objects in the direct context of attributes, is known as a "data context map." Fig. 1 represents a criteria to find a dream university, which is based on various criteria like good academics region, good athletic region, low tuition region and combined region [1]. The data context map has many advantages because it allows users to appreciate 1) similarity of data objects, 2) similarity of attributes within the scope of data objects, and 3) relationships between data objects and the attributes. This layout has also allowed various data regions, which are segmented and hence labelled based on the various locations of the attributes. And hence, this helps in the selection of tasks where users seek to identify one or more data objects.**

***Index Terms*—Multidimensional matrix, Segmented, Data context map**

I. Introduction

The data matrix is one of the most fundamental structures in data analytics. It is a M*N rectangular array of N variables and M samples. The N*N or M*M similarity matrix S is another frequently used structure derived from DM. So as to visualise DM, the current methods either focus on preserving the relations among the samples or on preserving the relations among the variables, but they are typically not capable of doing both of them. This is one of the major limitations with which one wishes to transform DM into a comprehensive map in which the acquired samples are accurately present in the context of variables. In the paper, the authors Cheng and Mueller [1] describe the new data and similarity matrix that overcome these deficiencies.

In order to illustrate the above-mentioned points, the authors consider an example in which a parent is looking for a university for their child, as shown in Fig. 1. And the factors to be considered for this are: academic score, athletics, tuition, teacher-to-student ratio, and many others. College Prowler is a popular website that allows users to navigate this parameter space by narrowing the search using slider bars and menu selections for each parameter.However, it is a tedious process. On the other hand, a visual expert would use interactive parallel coordinate plots, but it is difficult to imagine that a normal parent would engage in such an advanced interface. Also,

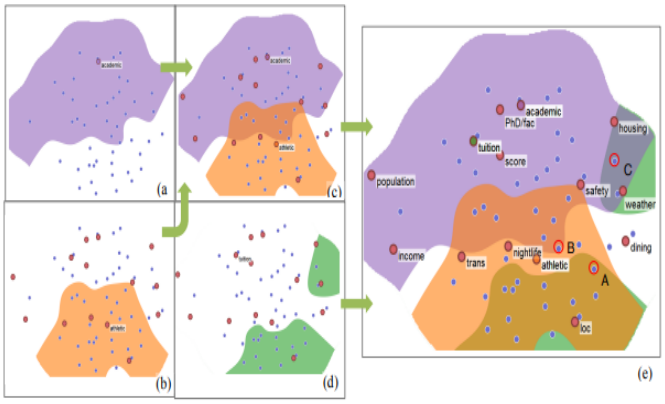

Fig. 1: The process to find a dream university. (a) good academics region. (b) good athletic region. (c) combined region by (a) and (b). (d) low tuition region. (e) combined region by region (c) and (d) Image courtesy: [1]

there are several other visualisation methods, such as interior layouts, but these can be easily found in the mainstream arena. On the other hand, the widespread familiarity of maps makes them a natural canvas to overview the landscape of universities in the context of various factors to be considered. And hence, parents could simply sit back and examine the illustration like an infographic, then decide on a school. Also, they could still use a filter to eliminate some schools from the map, but they would never lose sight of the big picture. There are several methods, like multidimensional scaling (MDS), self-organizing maps (SOM), locally linear embedding (LLE), t-distributed stochastic neighbour embedding (t-SNE), etc., that create 2-D map-like data layouts that are computed from the similarity matrix S of schools. These maps show similar schools as clusters and special schools as outliers, which is certainly useful, but parents will not be able to tell from the plot alone why some schools are special and others are clustered. Also, their ranking among athletics, tuition, etc. will not be known easily.

The school similarity matrix S could only hold attribute similarities. The maps would then allow a visual assessment of the grouping of the attributes. Hence, instead of finding that schools A and B are very similar or dissimilar in terms of their attributes, one would find that attributes C and D are correlated or not in this set of schools. As a result, parents may discover that the higher the academic score, the higher the tuition, and the more students per teacher, But if the parent is interested in smaller classes, then schools with lower academic scores might be a better choice. And hence, while such a plot is useful in explaining the relationships between the different features of the educational landscape, it still cannot allow the anxious parents to pick a specific school for their child, which

is what they need.

In the paper, the authors Cheng and Mueller [1] proposed a framework that overcomes the limitations and combines both of the similarity aspects derived from DM into a single comprehensive map, which the authors called the data context map. This method requires a non-trivial fusion of the two alternative similarity matrices S. With the help of this fused matrix, a mapping can be performed that allows the users to faithfully appreciate all three types of relationships in a single display: (1) patterns of the collection of samples; (2) patterns of the collection of attributes; and (3) relationships of the samples with the attributes. Also, the contextual mapping provides the information needed to add semantic labelling to the samples. and hence iso-contouring these regions to create decision boundaries by which one can easily recognise trade-offs among different samples, as it can be helpful in complex decision-making. The authors, Cheng and Mueller [1], have demonstrated these ways with the help of a few practical examples.

II. RELATED WORK

The visualisation of high-dimensional data on a 2D canvas mostly follows the three paradigms of projective data displays, interior displays, and space embeddings. But since the visualisation of high-dimensional data in 2D is inherently problematic, there is no method without drawbacks. Therefore, it is almost impossible to preserve all the variances of a high-dimensional point cloud in a 2D mapping. Hence the different methods that have been described, which offer different strengths and weaknesses.

A. Projective and interior displays

This method displays wrap-up data in a way to emphasise certain properties, such as locality or similarity. A projective display is a scatter matrix that displays the scatter plot. It reserves a scatter plot title for each variable and projects the data items into it. Such a division of data context into two variables makes appreciating the overall pertaining of all variables difficult. Furthermore, because the data points are located far away in high-dimensional space, they may project into similar 2D locations during the mapping operation. This further adds to the difficulties in recognising multivariate relationships.

The parallel coordinates and the star plot, which is their radial version, represent the variables as parallel or radial axes, respectively, as they map the data as polylines across the axes [1]. However, the clutter of polylines is likely to become a significant problem once the number of dimensions and data points increases. So in order to decrease the clutter of lines, the star coordinates arrange themselves in a radial fashion instead of constructing polylines; also, they plot the data points as a vector sum of the individual axis coordinates. A vector sum is an aggregation that maps the data to locations that are not unique. In other words, there are points that map to nearby locations but may not be close in high-dimensional space, and vice versa. In order to help users resolve these ambiguities, at least partially, an interactive interface is often provided that allows them to rotate and scale the data axes and uncover false neighbors.

There are several displays that are similar to spherical coordinates and have the same drawbacks. And hence they are called ”interior displays,” since they all lay out the variables as dimension anchors around a circle and map the data items as points inside them, given some weighting function that relates to the data points as different attribute strengths. Also, all of these displays are useful in what they have been designed to convey, that is, the relation of data points with respect to the attributes. as the mapping function does not involve the similarity of the data points and results in ambiguities. The current framework is completely different in that it maps the attributes not only in the periphery along a circle but also intersperses them throughout the data distribution, significantly reducing all mapping errors. It allows for the labelling of regions as well as the setting of decision boundaries.

B. Comparing the interior displays

The method of GBC can serve as a standard reference framework that describes most of the interior displays. The GBC plot uses the dimension values of an N-D point as weights in a weighted sum of the 2D locations so as to determine the point's location in the 2D polygon. Using these GBC plots, they also conducted a controlled experiment and compared them with the method proposed here. For this, they generated a test dataset comprised of a set of six 2-D Gaussian distributions. As proposed by the author, first they randomised the six 6-D centre vectors and then randomised 600 data points following these distributions. The dataset is visualised using parallel coordinates, assigning each Gaussian a unique color. In addition, the axes are also coloured such that each axis colour matches that of the cluster with the highest value for that dimension. The following results are shown: (a) standard GBC compared with (b) the optimised GBC plot; and (c) the method proposed in the paper that allows the attribute nodes to scatter among the samples. It is shown that the method proposed is more flexible and can preserve the pairwise distances well.

C. Embedded displays

The ambiguities that are present in the relationships between the data points are often overcome by embedding the high-dimensional space into the 2D canvas. Principal component analysis (PCA) finds the two eigenvectors that are associated with the largest variation in the data and then projects the data points into the plane that is spanned by these vectors. There are other methods that seek to create a mapping from high-dimensional to 2D space that optimises for some measure of data point similarity. MDS is aimed at preserving some distance metrics, such as Euclidian distance or pattern distance. There are other mappings, such as ISOMAP, LLE, SOM, t-SNE, LAMP, and PLP, that optimise for geodesic distance, distribution distance, locality, etc. Hence In these 2D embeddings, the viewer can appreciate neighbourhood relations and obtain a good overview of the space quickly. In case the users wish to see the relationships between both

attributes and data samples, then there are two separate maps that need to be created using the two alternative forms of the similarity matrix S as presented in the introduction, one for the samples and one for the attributes. However, these methods have the drawback of no longer maintaining any context with the attribute space because this information is typically not preserved in the mapping. The method described in this paper fuses the two alternative similarity matrices and so is able to create an embedding for which the relationships among samples, among attributes, and among the two of them are equally preserved. In practice, the authors Cheng and Mueller [1] use a dissimilarity matrix as well, because a similarity matrix is simply the opposite of dissimilarity.

D. Fused Displays

The work on fused displays is relatively uncommon [1]. One of the recent implementations, which is similar to the approach given by the authors Cheng and Mueller [1], has created a fused matrix of samples and attributes as it uses 2D layouts. There are the following differences between the two approaches: • Their framework is designed for categorical data. The numerical data are binned into regular intervals, which can be inaccurate. The approach proposed by the authors starts with numerical data by default as it transforms the categorical variables into numerical ones, by taking into account the pairwise distribution of relationships [1]. • They mostly use a linear projection approach, which is based on multiple correspondence analysis (MCA), to create the 2D mapping. The layouts are generated via numerical optimization, which can support a variety of constraints and also better preserve high-dimensional relationships. • They compute a diagram so as to divide the domain into value regions, which only account for the relationships among the attributes and their levels. The approach presented by the authors generates a set of general iso-contours computed from a continuous heat map of the data using adaptive kernel density estimation.

III. THEORY AND METHOD

In this method, the authors Cheng and Mueller [1] created a mapping in which all three types of relationships in DM are preserved: those among the samples, those among the attributes, and those mutually among the samples and attributes. Hence the notion of relationship can be a distance, such as Euclidian, or a similarity, such as Pearson's correlation, cosine, or pattern, or it can be some measure of significance, such as value or feature. Hence they combine all of these functions collectively into a distance metric, F, and note that, depending on the application, each of these relationships might be expressed in a different F. For instance, there is a similarity of attributes that might be measured by the correlation, while the proximity of samples might be gauged via the Euclidian distance. The author wished for a mapping that preserved this set of simultaneous constraints. And therefore, it calls for an optimization strategy based on a fused representation for all three types of relationships.

The authors, Cheng and Mueller [1], proposed the various steps of this pipeline in detail. The distance matrix, which is one for each of the three pairs, encodes the respective F. The fusion process hence merges each of these three matrices into a single distance matrix, emphasising certain constituents and equalising them, which is followed by a mapping to 2D using an optimization process.

A. Data Matrix

In the data matrix, the rows denote the data samples, whereas the columns denote the variables, and xij is the data value in the ith row and jth column. Without losing generality, it was assumed that the DM is normalised to [0, 1]. So depending on if the DM is row-wise or column-wise, they have two types of spaces: the data space D and the variable space V, respectively. All m data items (examples) are contained in the data space D.

Di = [xi1, xi2,.....xin] where i=(1,2,3.....m)

As a result, the n orthogonal attribute axes span it. A variable space V, on the other hand, contains all n data attributes:

Vj = [x1j, x2j,... xmj], j = (1, 2,... n). and thus is traversed by m orthogonal data item axes.Also, the data space D is the more familiar of the two, but there are many applications in which the samples can turn into attributes and vice versa, depending on the focus of the analytics. For instance, for a data matrix storing the results of a DNA experiment on multiple specimens, one of the research objectives might consider the genes expressed in the microarray to be the samples and the specimens to be the attributes.

B. The Composite Distance Matrix (CM)

In the next step, the process is to define the desired distance or similarity metric for each of the relationships. As shown in Fig 2. for mapping the similar items into closer proximity, they also need to use (1-correlation), (1-attribute value), etc., while the spatial distance metrics, such as Euclidian, can also be used as they are. The authors, Cheng and Mueller [1], have four different distance matrices: • DD to store the pairwise distance of the data items; • VV to store the pairwise distance of the attributes. • VD to store the pairwise distance of attributes to the data items; • DV to store the pairwise distance of the data items to attributes. A point in the space that is very well defined for the values of the variable that has the samples for the data columns of DM.

C. Fusion

To merge or fuse the two spaces, a set of transformations—scale, rotation, and translation—is required. For the time being, the authors, Cheng and Mueller [1], have also implemented only scaling. The four matrices that make up CM were not created equally. They have been calculated from vectors with different lengths of n or m, and they may have used it in different distance metrics (F). As observed, this inequality can lead to cases in which data samples and attributes do not mix well. And hence, the points due to the data samples and those due to the attributes may clump together into separate and disjoint communities. Thus, the

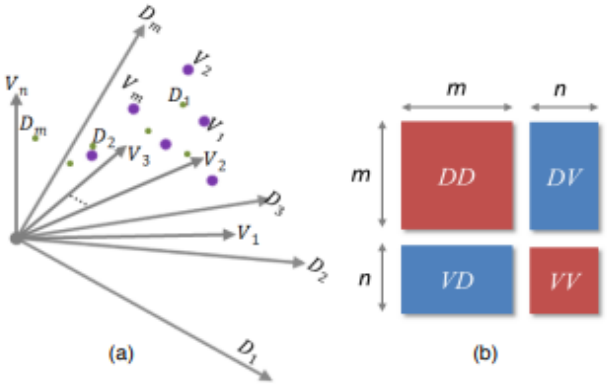

Fig. 2: (a) The fused space composed of D and V and (b) the composite distance matrix CM and the extents of its submatrices DD, DV, VD, and VV. Image courtesy: [2]

transformations are necessary to enlarge or shrink the data or variable spaces.

In order to mix the data and variable spaces well, there should be a balance between the differences of each of the four matrices. Making the four sub-matrices have equal means is one of the simplest ways to define or achieve this. And in this way, the two spaces have equal scale. There are various options to make the four distance matrices have the same mean as a linear, polynomial function. Also, a linear function has many advantages because it preserves the topology, which is presented in the paper, for which we can apply a linear weight adjustment for each submatrix.

D. Mapping

The joint map of samples and attribute points is produced once the composite distance matrix (CM) has been obtained. As opposed to a linear projection with PCA or biplots, the authors Cheng and Mueller [1] have chosen to employ an optimization technique for the map layout since it provides more flexibility in selecting the constraints guiding the layout, such as mixed distance functions, layout schedules, and mapping criteria. Many distance-preserving optimization methods are suitable for the needs of the authors. Locally optimal layouts are produced by LLE, but globally optimal layouts are produced by MDS-type schemes, which have grown in popularity recently since they offer a consistent picture of the data. Last but not least, t-SNE or linear discriminant analysis (LDA) excel at isolating specific clusters.

E. A First Example

The MDS layout for just the data samples using the Euclidian distance metric is the same as the MDS configuration for the entire dataset using the Euclidian distance metric. an MDS layout for the attributes using Pearson's correlation distance, an MDS layout using the entire CM matrix and weights set to not emphasise any CM submatrix, and a parallel coordinate display for this dataset with the axes coloured in accordance with the attribute nodes. For the DV and VD submatrices, the 1-value distance was employed.

The first thing they notice is that the CM-based MDS layout has done a good job of maintaining the clusters' original layout in the sample-only MDS displays. On the other hand, the characteristics' locations—while still mostly separated to account for variations in correlation—have changed and now more accurately reflect their relationships with the data clusters. As a result, it can be seen that the joining of the two spaces D and V involves more than simply the simple superposition of the two plots.

More precise observations include the following: (1) the red cluster clearly dominates the red attribute, and its dimension node is mapped into the red cluster's centre; (2) the green and brown clusters both have high values for the green attribute, so the node for the green attribute is mapped between these two clusters; (3) the same is true for the brown attribute and the red and brown data clusters; (4) the dark blue and black attributes have such strong correlations with the red and brown data clusters.

Upon closer examination, it appears that CM's layout gives little or no weight to the lower levels of the qualities. The distance measures they choose in this situation can help to explain this. The 1-value distance they choose for the DV and VD submatrices is the reason why the algorithm prefers to pick attribute placements around high values in the data clusters. A different distance would have resulted in different behaviour. The analyst's preference for highlighting certain features of the data will have a significant impact on this and other decisions, as well as their outcomes. The emphasis in this example was on extreme values.

IV. CONSTRUCTING THE DATA CONTEXT MAP (DCM)

The authors describe this section in order to provide more details on the map's construction and segmentation into regions with similar properties.

A. Populating the map

Since the weights of the CM submatrices can not only be altered during the MDS layout, but can also be applied to various MDS schedules for the samples and attribute points. This idea was used by the writers to create layouts with various priorities. To accomplish this, as shown in Fig 3 they need an iterative MDS algorithm. Algorithms for iterative MDS frequently do not update all points at once. Instead, they choose a subset of points that can move while others must stay put, either permanently after an initial configuration or alternately. The point sets may also be ephemeral and subject to change. The Glimmer MDS (G-MDS) algorithm is a particularly useful approach in this area.

Four MDS schedules are currently provided by the authors using this adaptable updating scheme: (1) Change both the variables and the data points at once (M-MDS); (2) map the variables first, then fix them and only map the data (VF-MDS); (3) map the data first, then fix them and only map the variables (DF-MDS); and (4) the user-defined order (U-MDS).

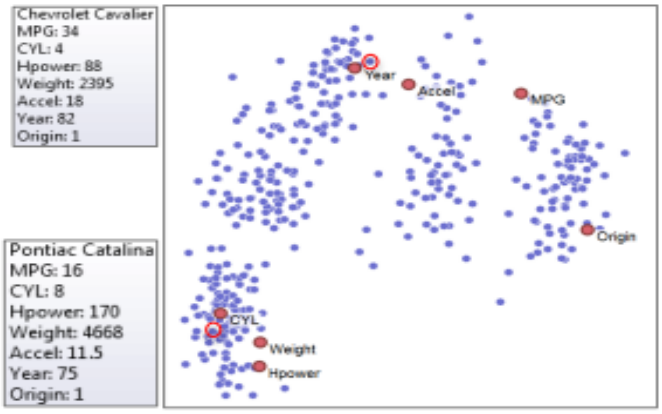

Fig. 3: The data context map for the car data. Image courtesy: [3]

B. Segmenting the Map

According to the authors Cheng and Mueller [1], the data context map as it has been presented thus far already allows attribute-informed selection of data objects. The evaluation of the various value ranges for combinations of attributes, however, was quite difficult. This would be easy if the map could be separated into distinct geographic zones that could then be recognised by their matching attribute value combinations. To do this, they require a continuous map representation. They have used adaptive kernel density estimation for this (AKDE). They then used AKDE to generate the property distance field. Constructing contour fields is a typical method for visualising distance fields. A closed range of attribute values results in a filled region between the two matching contours, and each attribute gives rise to a set of contours. The decision areas can be used in order to further this goal. The combination of these properties is then used to encode the entire area. An entirely segmented and self-labeled map is made to do this.

V. IMPACT OF THE PAPER

As per Google Scholar, the given paper is being cited by 75 other papers. After reviewing the various other papers cited in the given paper, the author concluded that several theories emerged.

According to the author of this paper, there are various nodes present in clustered social networks that cannot be easily assigned to any clusters. Hence, these nodes are found either at the interface between these clusters or at the boundaries. There is a strong need to identify these nodes because they are important in marketing applications such as voter targeting because the people represented by them are more likely to be affected by marketing campaigns than the nodes themselves. Hence, this identification task is not as well studied as the other networks. The problem of identifying interface and boundary nodes is being addressed in clustered social networks. And the presented solution is that the clusters might correspond to political parties in the political discussion networks. A common property of boundary and interface nodes is that they do not belong to any of the clusters. The people represented by these nodes are more interesting targets for marketing campaigns that aim to grow with existing clusters. Hence, the author presented two case studies in which the interactive approach turned out to be effective in identifying interface and boundary nodes. [2]

The paper's authors provided an in-depth analysis of two case studies of visual analytics in which both machine- and human-centric approaches were used for classification tasks. One such study is a new application for the classification of visual images, while the other is for facial expression classification. It is observed that in both studies, a human-centric approach produced better decision trees than a machine-centric approach. Also here, the authors estimated the Shannon entropy for the various alphabets that are featured in the two pipelines for machine- and human-centric approaches. [3]

In this paper, the author presents an empirical evaluation of the dimension reduction techniques from the perspective of perception in visual cluster analysis. Various dimensionality reduction techniques are identified from a literature point of view. Also, they are grouped into groups based on the various kinds of linearity and locality. The author also conducted a pre-study so as to determine the proper input parameters for each dimensionality reduction technique. Many interesting insights and results are provided in the given paper. [4]

VI. CONCLUSION

The authors, Cheng and Mueller [1], have discussed in the paper that the data context map is a framework and visual interface that enables a comprehensive layout for both data points and variables. They accomplish this by combining the two distance matrices, the data and attribute distance matrices. The authors also created an optimised layout that can be used for data-driven decision selection and decision problems that totally require a mindful balancing of trade-offs. As provided in the paper, there are several parameters for the experts to guide the layout for their goals, but they are not essential to producing usable results. Hence, casual users can easily use the pre-set weights, upload the data, generate the initial map, and interact with the value sliders. The future work in this area can be about how casual users actually get to do this.

REFERENCES

[1] Shenghui Cheng, Klaus Mueller, "*The Data Context Map: Fusing Data and Attributes into a Unified Display*," IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 22, NO. 1, JANUARY 2016.

[2] Shenghui Cheng, Joachim Giesen and Tianyi Huang, Philipp Lucas, and Klaus Mueller, "*Identifying the skeptics and the undecided through visual cluster analysis of local network geometry*".

[3] Gary K. L. Tam, Vivek Kothari, and Min Chen, "*An Analysis of Machine- and Human-Analytics in Classification*," IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017.

[4] Jiazhi Xia, Yuchen Zhang, Jie Song, Yang Chen, Yunhai Wang and Shixia L, "*Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study*," IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 28, NO. 1, JANUARY 2022.