Abstract

In typical data mining applications we often have large amounts of data at our disposal along with knowledge often available in quite different ways such as rules, cases, analytical models or correlations among variables. Many classical machine learning methods may result inadequate in this scenario because they seldom allow to make use of all the knowledge that we might have at hand. Visualization techniques, that have been used for a long time for data visualization, can also be used to visualize certain forms of knowledge, resulting in a more efficient data mining process. In this paper we present a unifying approach for knowledge visualization based on dimension reduction (DR) that allows to represent rules, cases, models and correlations on a low-dimensional visualization space in a consistent way.

1 Introduction

Problems arising in many emergent areas of the 21st century, such as gene expression analysis [13], content analysis of large document collections [15], complex process analysis [6][4], etc., often require the analysis of large volumes of high-dimensional data that the researcher has to correlate with any available knowledge about the problem obtained by alternative ways. This knowledge can be present in quite heterogeneous ways such as rules, physical or mathematical models relating two or more variables, correlations or case-based knowledge.

Integration and reuse of this knowledge along with information conveyed by the data has been a subject of very intensive research effort in the last decades. In particular, visualization techniques have received considerable attention for the analysis and understanding of large volumes of complex and high-dimensional data [12][23]. One approach for visual data analysis is the application of a dimension reduction (DR) mapping that allows to transform data from a high-dimensional space into a low-dimensional space (usually 2D, 3D for representation purposes) while preserving significant information related to the problem under analysis. This idea is not new; for instance, linear projection methods such as Principal Component Analysis (PCA) have been extensively used for data visualization and dimension reduction for decades. More general linear projection methods seeking the directions that maximize some degree of “interestingness” such as exploratory Projection Pursuit (PP) [8] and, more recently, Independent Component Analysis (ICA) [11][10][2] have received much attention in the literature. Nonlinear projection methods have also played a major role in the literature. Multidimensional scaling methods (MDS) like Sammon’s nonlinear mapping (NLM) [20], or more recent approaches like ISOMAP [22] and Locally Linear Embedding (LLE) [19] seek to provide a set of low-dimensional data whose mutual distances are similar to those of the original high-dimensional data, hence preserving the intrinsic structure of the data. Closely related, topology preserving methods like the self-organizing map (SOM) [14] or the generative topographic map (GTM) [1] allow to obtain mappings such that nearby points in the visualization space represent nearby points in the input space.

One key idea for the success of these methods is that all of them, to some extent, preserve information regarding the structure of the data in the transformation into a visualizable space, allowing us to find patterns related to available knowledge in a highly efficient way through visual exploration. A great part of this knowledge might be available in rather diverse ways, such as banks of rules, analytical models, known cases, or even in terms of correlations. In this paper we show how all these forms of knowledge can be given a geometrical interpretation in the high dimensional data space, suggesting a close relationship between knowledge and structure. Based on this idea we provide a global framework for representing both data and knowledge in a unified way through DR mappings.

This paper is organized as follows. In section 2 we show how different kinds of knowledge can be described in terms of scalar fields in the input space with a strong geometrical
2 The geometrical nature of knowledge

Finding knowledge in data has been often related as finding hidden patterns [9] or discovering natural structures [18]. Indeed, the kind of knowledge we usually seek in data is closely related to structure, and has a geometrical nature. Analytical models, for instance, consist of sets of relationships among variables that define manifolds in the data space. Rules are often expressed in terms of spatial constraints to the variables and geometrical-like operations such as intersections (AND), unions (OR) between regions in the data space. Similarly, correlations express geometrical relationships between variables. Most of the times, knowledge discovery in data (KDD) consists of finding matches between these geometrical models and the spatial distribution of data.

As we will show henceforth, these kinds of knowledge can be described as scalar fields in the data space $D$. Scalar fields implicitly define entities such as gradients, isosurfaces, etc. that suggest a geometrical conception of knowledge. This conception provides the basis for a unified way to visualize knowledge through dimension reduction.

2.1 Data sets

Let’s consider a set of data $\{x_i\}_{i=1,\ldots,N}$ under analysis. The information conveyed by these data can be modelled by a joint pdf $p(x)$ from which the $x_i$ are likely to come from. This joint pdf is indeed a geometrical model of the data distribution consisting of a scalar field in $D$ that describes the geometrical structure (both shape and density) of the input data.

2.2 Analytical models

Let’s define a model in the data space as a known relationship among the components $x_1, x_2, \cdots, x_n$ of $x$ – for convenience, written in implicit form–

$$f(x_1, x_2, \cdots, x_n) = 0$$

From this relationship a residual can be defined for each point $x \in D$ as

$$\varepsilon(x) = f(x_1, x_2, \cdots, x_n)$$

that gets close to zero when the model holds. This residual is indeed a scalar field that defines a manifold in $D$ for $\varepsilon(x) = 0$. Several models $f_i(x) = 0$ can be combined into a single model, leading to a vector of residuals

$$\varepsilon(x) = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m)^T$$

Any norm of this vector, $||\varepsilon(x)||$, defines itself a scalar field that reflects the joint fulfillment of the $m$ models.

2.3 Rules

Let’s consider a set of $r$ fuzzy rules (consider crisp rules as a special case) of the type

Rule $i$:
if $x_1$ is $A_{i1}$ and $\cdots$ and $x_n$ is $A_{in}$ then $C_1$ is $B_{i1}, \cdots, C_p$ is $B_{ip}$

The evaluation of the previous bank of fuzzy rules using any of the classical models for fuzzy inference (Mamdani[17], Sugeno [21], …) under a given choice of membership functions, defines $p$ mappings

$$\phi_{C_1}(x), \phi_{C_2}(x), \cdots, \phi_{C_p}(x)$$

from $\mathbb{R}^n$ to $\mathbb{R}$, taking a point $x \in \mathbb{R}^n$ and giving a value corresponding to defuzzified (crisp) values for consequent variables $C_1, C_2, \cdots, C_p$, respectively. As with pdfs, $\phi_{C_j}(x)$ represent scalar fields in the data space $D$.

2.4 Other forms of knowledge

In a similar way, many other kinds of knowledge can be expressed through scalar fields. We can compute new scalar fields using any kind of function, algorithm or procedure that yields a scalar value or measurable property for each point $x$ of $D$. For instance, in section 4.6 a local covariance matrix that describes the second order statistics of data in the neighborhood of $x$ is defined, allowing to provide local descriptions of correlations, or sparsity of data around $x$. Similarly, as we show in section 4.2, we can combine the variables of $x$ to obtain new variables $x_{n+i} = v_i(x_1, \cdots, x_n)$ that make physical sense to us.

3 Dimension Reduction Mapping

As seen, apparently different kinds of knowledge have a strong geometrical nature since all of them can be expressed through a scalar field in $D$. Finding patterns from data in knowledge discovery can thus be regarded as finding matches between these scalar fields and the structure of data. In other words, we seek to validate our knowledge by evaluating the associated scalar fields within the support of
the data, represented by regions where \( p(x) \) is significantly nonzero.

Using DR methods we can capture the low dimensional structure of data to get a map that represents the support of the input data without loss of significant information. Through this map we can validate our knowledge in the domain of the data. In general, DR can be defined in terms of a mapping

\[
\varphi : D \to V
\]

where \( D \subseteq \mathbb{R}^n \) is the data space composed of a large number of variables, usually meaningful, and \( V \subseteq \mathbb{R}^p \) is the visualization space, which contains a much lower number of variables—typically, two—that usually lack physical meaning but allow graphical interpretation. For our purposes, we will additionally require that an inverse mapping exists

\[
\varphi^{-1} : V \to D
\]

such that

\[
\varphi(\varphi^{-1}(u)) = u, \quad \forall u \in V
\]

and that provides a good approximation for points \( x_i \) of the input dataset

\[
\varphi^{-1}(\varphi(x_i)) \approx x_i
\]

These concepts are shown in fig. 1. Some of the most popular DR techniques, like SOM or PCA, define mappings which fulfill the self-consistency property (7). Other DR mappings, however, especially those obtained through interpolation of discrete DR methods—may not strictly fulfill (7), but define continuous mappings that closely approach it, being good enough for visualization purposes.

**Figure 1. Dimension reduction mapping.**

### 4 DR-Based Knowledge Visualization Methods

From equations (5)-(8), DR mappings implicitly define a manifold \( M \equiv \varphi^{-1}(V) \) in the data space \( D \), with the same intrinsic dimension of \( V \), but whose geometry approximates that of the data set under analysis. The points \( x^u = \varphi^{-1}(u) \) pertaining to \( M \subseteq D \) are therefore good representatives of the input data set and, in addition, have one and only one image \( u \in V \). This allows to represent any meaningful scalar property defined in \( D \) by evaluating it for every point \( x^u \) of the manifold \( M \) and displaying its value in \( V \) at position \( u \) by using a color scale. This defines a color pattern in \( V \) that will be referred to as plane or map. We will make use of this idea hereupon in this section to show how different forms of knowledge can be represented in \( V \).

### 4.1 Component planes

A great deal of prior knowledge is often related to the individual components \( x_1, \ldots, x_n \) of \( x \) just because they are meaningful. Component planes have been used long ago with Self-Organizing Maps [14] to provide a picture of how variables—components of vector \( x \)—behave in a given data set. They are formed by representing the value of each of the components \( x^u_1, x^u_2, \ldots, x^u_n \) of all points \( x^u \in D \), thus leading to \( n \) possible planes. Component planes have shown to be very powerful—as well as simple—representations of multivariate data in many engineering applications [16].

### 4.2 Virtual planes

Sometimes new variables with physical meaning for us can be obtained in a simple way from other related variables of \( x \) by means of functional relationships

\[
x_{n+k} = v_k(x) = v_k(x_1, x_2, \ldots, x_n)
\]

for \( k = 1, \ldots, m \). Thus, for instance, in an electrical system one can obtain instantaneous power, \( P(t) \) from instantaneous current, \( I(t) \), and voltage, \( V(t) \), as \( P(t) = I(t)V(t) \). Functions \( v_k \) stem from our physical prior knowledge about the system expressed through equations/models which relate other concepts. According to this, new planes can be represented by applying the functions \( v_k \) to the points of \( M \) and visualizing them in \( V \), in exactly the same manner as usual component planes

\[
x_{n+k}^u = v_k(\varphi^{-1}(u))
\]

where \( u \) represents a point of \( V \).

### 4.3 Density maps (cases)

Knowledge is often given in terms of cases consisting of data sets, \( L = \{ x_i \in \mathbb{R}^n \}_{i=1, \ldots, N_L} \), from which a certain amount of prior knowledge is available. As seen in section 2, the structure of this data set can be described by its joint pdf, \( p_L(x) \). This pdf actually embodies all domain
knowledge we may have developed regarding to the dataset (case). Thus, a density map showing up those regions that represent the probability mass of the subset $L$, provides a visual metaphor that we can associate to our knowledge about $L$. One way to do that is to evaluate –e.g. using a Parzen estimation– the joint pdf of the data subset for all points $\varphi^{-1}(u)$ of the DR manifold in $D$ and visualize them in $V$.

$$p_L(u) = \frac{1}{N_L} \sum_{i=1}^{N_L} \exp \left( -\frac{\|x_i - \varphi^{-1}(u)\|^2}{2\sigma_i^2} \right), \quad u \in V$$ (11)

This idea is described in figure 2, where the DR manifold $M$ and a data subset in $D$ are shown on the left and the corresponding density map in $V$ is shown on the right.

**Figure 2. Density Map.**

### 4.4 Model maps

Prior knowledge may also be available in terms of analytical models which describe relationships among the data variables. This knowledge can be used to obtain the regions in the visualization space that agree with known models.

Let’s consider a set of $m$ models of the data expressed in terms of implicit functional relationships between the involved variables

$$f_k(x_1, x_2, \cdots, x_n) = 0 \quad k = 1, \cdots, m$$ (12)

where $x_1, x_2, \cdots, x_n$ represent the components of a data vector $x$ from $D$. These equations may come from Physics or from experience, and usually convey a lot of prior knowledge about the underlying phenomenon or process being analyzed. On the other hand, they represent high-dimensional geometrical entities –manifolds– in the data space $D$ whose intersections with the data support can be represented in $V$, showing up the portions of the data that fulfill the model.

In effect, for each point $u$ from the visualization space $V$, a point $x^u = \varphi^{-1}(u)$ can be found in $D$ which is the “best” representative of those points in the input data set that are projected in $u$. Applying the model to the components of $x^u$, $m$ residuals can be obtained

$$\varepsilon_k(u) = f_k(\varphi^{-1}(u)) = f_k(x_1^u, \cdots, x_n^u)$$ (13)

for $k = 1, \cdots, m$. Each residual $\varepsilon_k(u)$ represents a measure of how the points $\varphi^{-1}(u)$ of the DR manifold in $D$ fit the model $f_k(\cdot) = 0$.

Residuals $\varepsilon_1, \cdots, \varepsilon_m$ can be visualized in $V$ through the so called model maps. The $k$-th model map is defined by the value of $\varepsilon_k(u)$ at each position $u$ of $V$ leading to $k$ color maps, where regions with near-zero values of $\varepsilon_k$ represent data points $x_i \subset D$ that are well represented by the model $k$. The idea of model maps is illustrated in fig. 3.

**Figure 3. Model $z = 2x + 0.5y - 1$ and data manifold induced by $\varphi$ in the data space $D$ –left–, and in the visualization space $V$ –right–.**

### 4.5 Fuzzy maps

The same approach can be used to exploit symbolic knowledge about the data by displaying automatically the regions corresponding to different classes inside data $[3][7]$. As shown in 2.3, if some knowledge about our data is available in terms of a bank of fuzzy rules, a set of $p$ nonlinear mappings

$$\phi_{C_1}(x), \phi_{C_2}(x), \cdots, \phi_{C_p}(x)$$ (14)

can be defined showing the degree of fulfillment of consequents $C_1, \cdots, C_p$. Each mapping $\phi_{C_k}$ can be applied to any vector $x \subset D$ to obtain the membership of consequent $C_k$. In particular, it can be applied to points $x^u$ of the DR manifold in $D$ and visualized in $V$ at point $u$. This leads to $p$ fuzzy maps, $c_1, \cdots, c_p$ showing the degrees of membership to classes $C_1, \cdots, C_p$ for every point $u$ in $V$

$$c_k(u) = \phi_k(\varphi^{-1}(u)) = \phi_k(x^u) \quad k = 1, \cdots, p$$ (15)

The $k$-th fuzzy map is thus defined by means of grey or color values that express $c_k(u)$ at each position $u$ in the visualization space $V$.

### 4.6 Correlation maps

Correlation analysis is concerned with finding how components $x_1, \cdots, x_p$ of the sample data vectors $\{x_i\}_{i=1,\cdots,N}$...
are mutually related for the whole data set. However, in many cases, data variables can be correlated in different ways for different regions of the data space. This is the case, for instance, of multimodal or nonlinear processes, which behave locally in different ways depending on the working point. We shall see hereupon how we can give a local description of correlations, that are meaningful in $D$ and can be visualized in $V$ by following the same principles.

**Local Correlations.** Let’s define a neighborhood function $w_i(u) = e^{-\frac{1}{2}\|x_i - \varphi^{-1}(u)\|^2/\sigma^2}$, describing the degree of proximity of sample $x_i$ with respect to $\varphi^{-1}(u)$ in $D$. We define [5] the local mean vector $\mathbf{m}(u)$ and the local covariance matrix $\mathbf{C}(u) = (c_{ij})$ associated to a point $u$ in the visualization space $V$ as

$$
\mathbf{m}(u) = \frac{\sum_i x_i \cdot w_i(u)}{\sum_i w_i(u)} \tag{16}
$$

$$
\mathbf{C}(u) = \frac{\sum_i (x_i - \mathbf{m}(u))(x_i - \mathbf{m}(u))^T \cdot w_i(u)}{\sum_i w_i(u)} \tag{17}
$$

where the width factor $\sigma$ is a design parameter related to the degree of locality to be taken into account, allowing to establish a tradeoff between global and local correlations. In a straightforward manner, a local correlation matrix can be derived from $\mathbf{C}(u)$ as

$$
\mathbf{R}(u) = (r_{ij}) \text{ where, } r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}} \tag{18}
$$

that provides a much more efficient description of correlations restricted into the interval $[-1, +1]$. Both the covariance matrix $\mathbf{C}(u)$ and correlation matrix $\mathbf{R}(u)$ are defined for each point $u$ of $V$ and provide a local description of second order statistical features of data in $D$ lying in the vicinity of $\varphi^{-1}(u)$. In the same way as in previous sections, the elements of both matrices $c_{ij}(u)$, $r_{ij}(u)$, the eigenvalues of the covariance matrix $\lambda_i(u)$ or the components of the principal vectors $s_i(u)$ can also be represented using planes.

### 5 Example.

A simple example is illustrated in fig. 4, where data points show up different local correlation patterns (the ellipses represent local covariances). For variables $x$ and $y$, points A show up mild positive correlations, points B have a strong positive local correlation and points C are inversely correlated. A SOM based DR mapping was computed. In fig. 5, correlation maps for components $r_{xx}, r_{xy}, r_{yx}, r_{yy}$ are displayed that clearly show up different correlations between $x$ and $y$ depending on the cluster A (light gray zones=mild positive correlations), B (light zones=positive correlations), or C (dark zones=negative correlations).

**Figure 4. Local covariances.**

**Figure 5. Local correlation planes.**

This representation is consistent with others such as fuzzy or model maps—see fig. 6—. Model $y = -x + 2.5$, with a negative slope, holds (dark tones=small residuals) in a region of high negative correlations including the whole set C. We also show the fuzzy map “$x$ is (VL or L or M) and $y$ is (L or M)”—see membership functions in fig. 4—that covers the whole set A (light tones=high degree of fulfillment). As seen, all information provided by maps of figs. 4, 5 and 6 is consistent.

**Figure 6. Fuzzy and model maps.**
6 Conclusion

In this paper we give a unifying approach for domain knowledge visualization using dimension reduction methods. We have shown how rather different kinds of knowledge such as rules, cases, models or correlations, can be expressed in terms of scalar fields in $D$ susceptible of being evaluated within the support of the input data. This is equivalent to particularize our domain knowledge to the available evidences provided by data.

Dimension reduction mappings, on the other hand, allow to provide an approximate –but often sufficient for knowledge discovery– representation of the data distribution by capturing its low dimensional structure and projecting it into a low dimensional space, $V$, that represents the whole domain of the data.

We show in this paper how the one-to-one correspondence between $V$ and a manifold $M$ that approximates the data layout allows to represent those forms of knowledge that can be expressed in terms of scalar fields, by means of color maps –planes– allowing us to relate them all in a unified and consistent way. This provides a visual approach for integrated data and knowledge exploratory analysis that can greatly enhance the data mining process.

7 Acknowledgements

Part of this work has been done under the research project ref. DPI2002-01599 financed by Ministerio de Ciencia y Tecnología and FEDER funds.

References


