A Multi-step Approach for Scientific Data Exploration

A. Ciaramella, F. Napolitano, G. Raiconi, R. Tagliaferri
DMI, University of Salerno, I-84084, via Ponte don Melillo, Fisciano (SA), Italy
{aciaramella,napolitano,graiconi,rtagliaferri}@unisa.it

G. Longo, A. Staiano
Department of Physical Sciences, University of Naples “Federico II”, I-80136, via Cintia 6, Napoli, Italy
{longo,staiano}@na.infn.it

In the modern age, the technological development allows the acquisition of more and more complex scientific data and today data analysis systems need advanced methodologies in order to mine and discover useful information. Aim of this work, therefore, is to propose an interactive multi-step approach for interactive data exploration. With data exploration we refer here, specifically, to clustering and visualization of complex scientific data. The data analysis process permits to explore and study raw datasets following, roughly, four steps: pre-processing, pre-clustering, agglomerative clustering, visual and interactive data exploration. One of the main advantages of the proposed approach is to divide the clustering phase in two steps: the first one is used to build an initial rough partition of the data in order to reduce the complexity of the problem. In the second step these clusters are presented to the user which searches interactively for structures in the data by means of a hierarchical agglomerative clustering approach. The proposed method has been developed in the Matlab computing environment and the implementation result is a user-friendly software tool, namely AstroNeural, which offers several clustering, visualization and data interaction techniques.

Key Words: Data Mining, Preprocessing, Clustering, Data Visualization.

1. INTRODUCTION

In these last years the field of Knowledge Discovery in Databases (KDD) is becoming of great importance for several fields of research. In fact, an explosive growth in the quantity, quality and accessibility of data which is currently experienced in all fields of science and human endeavour, has triggered the search for a new generation of computational theories and tools. They are capable of assisting humans in the extraction of useful information (knowledge) from huge amounts of distributed and heterogeneous data. At the core of the process there is the application of specific data mining methods for pattern discovery and extraction: in genetics, for example, several data mining approaches are proposed to analyse catalogues obtained from genome sequencing projects [1] [3] [7] [8] [5].

In this work we propose an interactive scientific data exploration tool that permits to accomplish both clustering/labelling and visualization tasks. The approach is mainly based on a hierarchy of two algorithms: a Probabilistic Principal Surfaces model [1] [2] constitutes the first clustering level, while the second one is constituted by an agglomerative method based on both Fisher and Negentropy information [1] [3] (NEC approach in the following).

Data visualization play an important role in extracting useful information from large quantities of data. A formidable pattern detection tool is ensured by human interaction and inspection. Unfortunately, the human eye needs low-dimensional data representations, usually two or three-dimensions, even though in real world applications data lies in very high dimensions. Moreover, quite simple relationships can seem very obscure when data are presented in a tabular form whereas they are often very easy to understand by visual inspection.

From the visualisation point of view, our goal is to enable the user to:

- Project and visualize data on a spherical surface (which provide a useful continuous manifold which can be rotated and manipulated in several ways);
• Perform deeper studies on the data by localizing regions of interests and interacting with the data itself: for example, choosing the points of interest, visualising their neighbours and similar points and printing of all related information etc.;
• Label the data choosing the classes found by the NEC approach;
• Visualise the labelled data in a two-dimensional space by using a Robinson projection and/or a Multi-Dimensional Scaling (MDS) approach;
• Select data to analyse.

The paper is organized as follows. In Section 2, we introduce the pre-processing problem. In Section 3, we detail the Probabilistic Principal Surfaces technique that represents the pre-clustering phase of the overall process. In Section 4, the agglomerative clustering and the visualization approaches are presented, in detail, Section 4.1 shows the agglomerative clustering based on Fisher and Negentropy information. In this section we also show how the hierarchical agglomerative tree obtained by the NEC can be used to interact with the clustering results. Moreover, the interactive data analysis and data exploration phases (Section 4.3 and Section 4.4, respectively) are shown by using the layouts of the developed tool. Section 4.5 covers the case of the availability of prior knowledge and the way to properly exploit it. Finally, in Section 4.6, the analysis of the computed clusters is detailed.

2. PRE-PROCESSING

Sometimes before to apply any clustering or whatever data mining technique, it is useful and also necessary to pre-process the data. Most real-world datasets, in fact, suffer of two major problems: noise and high dimensionality. These could derive from the method (or device) used to gain the data or from the nature of the data itself. However, in both cases a dataset could be intractable: on one hand, from the point of view of the accuracy of the results (high noise) and on the other hand, for the high computational cost (exponentially growing with dimensionality). The proposed tool includes different pre-processing methods, from data normalization to a non-linear robust PCA feature extraction.

3. PRE-CLUSTERING

In this phase the dataset is divided into different clusters, prevalently - but not necessarily - by using unsupervised techniques. This step allows the subsequent ones (which imply human direct interaction) to deal with a number of objects that can be smaller than the original ones by several orders of magnitude. At this aim, the Probabilistic Principal Surfaces is used to obtain a rough set of clusters that can be analysed and agglomerate by using a hierarchical agglomerative clustering (see Section 5). It is worth noting, however, that the proposed approach is independent from a particular pre-clustering technique and it could also be applied without pre-clustering or on pre-clustered datasets.

3.1 Probabilistic Principal Surfaces

Probabilistic Principal Surfaces (PPS) ([9,1,3]) are a non-linear extension of principal components, in that each node on the PPS is the average of all data points that project near/onto it. From a theoretical standpoint, the PPS is a generalization of the Generative Topographic Mapping (GTM) ([10]), which can be seen as a parametric alternative to Self Organizing Maps or SOM ([11]). Advantages of PPS include its parametric and flexible formulation for any geometry/topology in any dimension, guaranteed convergence (indeed the PPS training is accomplished through the Expectation-Maximization algorithm). A PPS is governed by its latent topology and owing to its flexibility, a variety of PPS topologies can be created, for example as a regular grid of a 3D sphere. A sphere is finite, unbounded and symmetric, with all the nodes distributed at the edge, and it is therefore suitable for emulating the sparseness and peripheral property of high-D data. Furthermore, the sphere topology can be easily comprehended by humans and thereby used for visualizing high-D data. Formally, PPS defines a non-linear, parametric mapping \( y(x;W) \) from a \( Q \)-dimensional latent space \( (x \in R^Q) \) to a \( D \)-dimensional data space \( (t \in R^D) \), where usually \( Q < D \).

The (continuous and differentiable) function \( y(x;W) \) maps each of the \( M \) points in the latent space (where \( M \) is the number of latent variables which is fixed \( a \ priori \)) to the correspondent point into the data space. Since the latent space is \( Q \)-dimensional, these points will be confined to a \( Q \)-dimensional manifold non-linearly embedded into the \( D \)-dimensional data space. We mark that the PPS approach builds a constrained mixture of Gaussians. Moreover if \( Q = 3 \) is chosen, a spherical manifold ([9,1,3]) can be constructed using a PPS with nodes arranged regularly on the surface of a sphere in \( R^3 \) latent space, with the latent basis functions evenly distributed on the sphere at a lower density. After a PPS
model is fitted to the data, several visualization possibilities are available like the projection of the data into the latent space as points onto a sphere. Having projected the data into the latent sphere, a typical task performed by most data analyzers is the localization of the most interesting data points, for instance the ones lying far away from more dense areas (outliers), or those lying in the overlapping regions between clusters, and to investigate their characteristics by linking the data points on the sphere with their position in the original data set.

Furthermore, the latent variable responsibilities can be plotted on the sphere in order to obtain the data probability density function visualization. Another advantage of the 3D sphere representation is that, unlike 2D plot, it is possible to observe the distribution of the different clusters on the sphere and observe which clusters are similar to each other as the dense regions that are in close proximity. All these advanced visualization options and successful applications to other research fields (i.e., in astrophysics) have been discussed in [1].

4. INTERACTIVE AGGLOMERATIVE CLUSTERING

After the application of the pre-clustering phase an interactive agglomerative methodology is applied to obtain the final partitioning of the data. Since rarely the number of clusters is known in advance, we adopt an interactive agglomerative technique, performed on the clusters obtained in the pre-clustering phase, where the user can choose different partitioning of the data on the base of a threshold value. The result is a dendrogram in which each leaf is associated with a cluster from the pre-clustering phase. The clustering technique is based on the Negentropy distance which can be computed directly between the clusters of the pre-clustering and that has several interesting properties [3,1], nonetheless this step could employ any hierarchical clustering technique since the clusters linkage can be performed using any metric.

4.1 Negentropy Based Clustering

The NEgentropy based Clustering (NEC) uses both Fisher’s and Negentropy information to agglomerate the clusters found in the pre-clustering phase. We underline that the approach we are describing is based on the one introduced in [3,1]. Such authors proposed a hierarchical agglomerative clustering where the optimal number of clusters is decided by analysing the plateaus obtained by varying an agglomerative threshold. We also stress that the most natural representation of a hierarchical agglomerative clustering is obtained by using the corresponding dendrogram, which shows how the samples are grouped [12]. In this paper the optimal number of clusters is defined using the dendrogram information.

On one hand we note that the Fisher’s linear discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space [13]. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. The Fisher criterion for two classes is given by

\[ J_F(w) = \frac{w^T S_B w}{w^T S_w w} \]  

(1)

where \( S_B \) is the between-class covariance matrix and is \( S_w \) is the total within-class covariance matrix. From equation 1 differentiating with respect to \( w \) we find the direction where \( J_F(w) \) is maximized.

On the other hand, the definition of Negentropy \( J_N \) is given by

\[ J_N(x) = H(x_{\text{Gauss}}) - H(x) \]

(2)

where \( x_{\text{Gauss}} \) is a Gaussian random vector of the same covariance matrix as \( x \) and \( H(.) \) is the differential entropy. Negentropy can also be interpreted as a measure of non-Gaussianity [14]. The classic method of approximate Negentropy is using higher-cumulants, through the polynomial density expansions. However, such cumulant-based methods sometimes provide a rather poor approximation of the entropy. A special approximation is obtained if one uses two functions \( G^1 \) and \( G^2 \), which are chosen so that \( G^1 \) is odd and \( G^2 \) is even. Such a system of two functions can measure the two most important features of non-Gaussian 1-D distributions. The odd function measures the asymmetry, and the even function measures
the dimension of bimodality vs. peak at zero, closely related to sub- vs. super-Gaussianity. Then the Negentropy approximation of equation 2 is:

\[ J_N(x) \propto k_1 E\{G^1(x)\}^2 + k_2 (E\{G^2(x)\} - E\{G^2(v)\})^2 \]  

(3)

where \( v \) is a Gaussian variable of zero mean and unit variance (i.e. standardized), the variable \( x \) is assumed to have also zero mean and unit variance and \( k_1 \) and \( k_2 \) are positive constants. We note that choosing the functions \( G^i \) that do not grow too fast, one obtains more robust estimators. In this way we obtain approximations of Negentropy that give a very good compromise between the properties of the two classic non-Gaussianity measures given by kurtosis and skewness \([3,1,14]\). They are conceptually simple, fast to compute, yet have appealing statistical properties, especially robustness. We have to note that several methods to accomplish Independent Component Analysis are based on entropy information. At this point we remark that our aim is to agglomerate by an unsupervised method the clusters (regions) that are found by a clustering approach. The information, that we call \( J_{NEC} \), used to merge two clusters is based both on the Fisher’s discriminant and on the Negentropy (NEC approach \([3,1]\)):

\[ J_{NEC}(X) = \alpha_F J_F(w) + \alpha_N J_N(X) \]  

(3)

where \( \alpha_F \) and \( \alpha_N \) are two defined (normalizing) constants and \( w \) is the Fisher’s direction. At this point using this information we apply the agglomerative hierarchical clustering approach and extract the dendrogram.

Algorithm 1 NEC: Agglomerative Hierarchical Clustering

Begin initialize \( c = \hat{c}, D_i \leftarrow X_i, i = 1, \ldots, c \)

DO \( \hat{c} \leftarrow \hat{c} - 1 \)

find nearest clusters, say:

calculate the Fisher’s direction between \( D_i \) and \( D_j \) and project the data on it

calculate the JNEC information and merge clusters \( D_i \) and \( D_j \) with lowest \( J_{NEC} \) information

UNTIL \( c = \hat{c} \)

return \( c \) new clusters

End

The NEC algorithm is described in Algorithm 1. To show the agglomerative hierarchical clustering process we describe the results on a synthetic data set. The data set that we consider is composed by two 2-dimensional classes with a complex distribution (see Figure 1). In Figure 1a we plot the data. In Figure 1b we show the clusters obtained using the PPS approach and in Figure 1c the obtained NEC dendrogram is plotted. We note that focusing our attention on the dendrogram we clear can define two distinct regions obtaining the clusters in Figure 2d.
4.2 Interactive clustering

The dendrogram plotted by our tool allows an interactive analysis. In fact, a dashed line, representing the current threshold, is shown on it and the corresponding partitioning is made evident by the use of different colours for the leaves of the dendrogram belonging to different clusters (see Figure 2, right). Each colour, in turn, is shaded from dark to bright according to the stability of the correspondent pre-cluster inside its cluster: the darker is the leaf, the sooner it will become an outlier when lowering the threshold (see Figure 2, right).

In fact, the threshold can be directly dragged on the dendrogram while the colours corresponding to the new partitioning are updated in real time. Also, all currently open visualizations (see next section) are updated accordingly.

Further information about the threshold is also shown: the current threshold value, the number of clusters, the number of outliers etc. Moreover, a menu containing all the possible clustering results (which are mainly characterized by the number of clusters and outliers it builds) is provided, allowing to directly choose the corresponding threshold value.

We stress that the key point of this clustering approach is that the user can arrange together great amounts of data operating on a human-tractable amount of information.
4.3 Data visualisations and exploration

As shown in previous sections, PPS, besides clustering, offer a convenient projection and visualization method. Other classical methods, such as Multi-Dimensional Scaling (MDS) [12] can provide visual aids about the distribution of the data. This is the process of converting a set of pair wise dissimilarities for a set of points into a set of coordinates for the points.

Often a far more significant visualisation can be obtained by merging the information given by such visualisations and the cluster information provided by the clustering phase. In particular, each of the points in the original data space can be associated with the colour of the cluster it belongs, which is provided by the dendrogram coloring. Even better, the shaded (pre-cluster) colour can be also used. After the association is obtained each visualization method can be improved adding to it the clustering colour.

We stress that our tool does this in real time giving immediate feedback to the user about the clustering he is considering.

This technique can be applied to projected dataset in 2D or 3D. In particular we have considered four kinds of visualizations:

- 2D Multidimensional Scaling
- 3D Multidimensional Scaling
- Spherical Probabilistic Principal Surfaces
- Robinson map projection of Spherical Probabilistic Principal Surfaces [6]

These visualizations can be shown contemporarily together with the dendrogram and updated in real time while acting on its threshold (see Figure 3). Our tool also allows to select sub-trees of the dendrogram and to trace back the involved data points simultaneously on the other visualizations. Also subsets of points can be free-hand selected on the 2D visualizations to save portions of the dataset identified as interesting by visual inspection. In fact in Figure 3 we show how we can use the 2-dimensional MDS visualization to select arbitrary data points for further analysis (see Figure 4).
4.4 Visualization of prior knowledge

It often happens that some or all the objects of a dataset are known to belong to some classes. This information can be used in at least two ways, to:

- Validate a clustering result.
- Infer new knowledge.

The validation of a clustering can be obtained comparing the prior knowledge and the knowledge obtained from the clustering itself. We note that this can give a confidence degree about the use of the same partitioning with data on which no prior knowledge is available.

The prior knowledge can also be used to produce new knowledge inferred by the presence or absence of objects of a certain class in a certain cluster. Different hypotheses can be made depending on the relations between the prior knowledge and the features used to cluster the objects.

Prior knowledge can be visually shown using different forms of points in the dataset belonging to different classes (see Figures 3 and 4, e.g. circles, squares, triangles). This permits to visualize, on all the visualization methods described, distance and cluster information together with the prior knowledge. Our tool allows the use of different sets of prior knowledge information and it permits to easy switch from one set to another.
5. CLUSTER ANALYSIS

Once the final partitioning of the data is obtained, analytic measures can be employed to assess its meaning. Typical measures are the mean and variance of the features of all objects in a cluster. In this process a well characterized cluster is, certainly, one with low variance, while clearly separated clusters must also have clear different mean behaviours.

Our tool is able to summarize all this information, including a mean and variance plot, for the objects falling in any subset of the dendrogram, or for the clusters defined by the threshold.

6. CONCLUSIONS

In this work we presented a multi-step approach to data exploration allowing the reduction of information contained in big datasets to a human-tractable size. We introduce the two main approaches of the overall model: PPS method for pre-clustering and the NEC hierarchical agglomerative clustering. We showed how a user is supported during the interactive part of the process with different kinds of data visualizations and how he can work with the data projected in a reduced space. We also showed how the results of the process are analytically checked to assure the significance of the clustering. Our software implementation is realized under the Matlab computing environment which provides the interactive environment for the exploration of data and the results.

REFERENCES.


