MedVir: An Interactive Representation System of Multidimensional Medical Data Applied to Traumatic Brain Injury's Rehabilitation Prediction

Santiago Gonzalez[1], Antonio Gracia[1], Pilar Herrero[1], Nazareth Castellanos[2], and Nuria Paul[3]

[1] Computer Science School, Universidad Politécnica de Madrid, Madrid, Spain
{sgonzalez,pherrero}@fi.upm.es, antonio.gracia@upm.es
[2] Center for Biomedical Technology, Universidad Politécnica de Madrid, Spain
nazareth@pluri.ucm.es
[3] Department of Basic Psychology I, Complutense University of Madrid, Spain
napaul@med.ucm.es

Abstract. Clinicians could model the brain injury of a patient through his brain activity. However, how this model is defined and how it changes when the patient is recovering are questions yet unanswered. In this paper, the use of MedVir framework is proposed with the aim of answering these questions. Based on complex data mining techniques, this provides not only the differentiation between TBI patients and control subjects (with a 72% of accuracy using 0.632 Bootstrap validation), but also the ability to detect whether a patient may recover or not, and all of that in a quick and easy way through a visualization technique which allows interaction.

1 Introduction

The possibility of detecting if a patient with a traumatic brain injury (TBI) can be rehabilitated by means of a treatment is not an easy work, however it is interesting. This is because it would allow to adjust and personalize treatments (in time and economy) of TBI patients to the needs of each one. One of the possibilities of evaluating the impact of brain injury is using MagnetoEncephaloGraphic (MEG) recordings through the obtaining of the functional connectivity patterns. These recordings are performed during several minutes of brain activity per individual (that is, time series of 148 sensors included in MEG machine).

The data generated by MEG are very complex (multidimensional and multivariate) and require much time for analysis. Nowadays, the idea of getting a prediction of the evolution of a TBI patient is out of the reach of clinicians, even without taking into account the analysis time. But if the data gathering process through MEG happens repeatedly, then is further complicated, so it is necessary to carry out an analysis mechanism that allows to draw conclusions (and extract new knowledge) in a quick and easy way.

In this work we were aimed to design a 3D visual interface for medical analysis easy to be used by clinicians. MedVir is a robust and powerful 3D visual interface to analyze, in this case, MEG data. After several stages, MedVir represents the information in two and three dimensions. Furthermore, the interface allows the experts to interact with the data in order to provide a more exhaustive analysis in the shortest time possible.

The paper is organized as follows: in section 2 a short overview of related work is presented. Section 3 presents the MedVir framework. Section 4 describes the TBI data obtaining and the experiments carried out. Finally, the conclusions and future lines (section 5) are reported.

2 Related Work

Nowadays, from the Data Mining point of view, there is no researches about TBI analysis through MEG recordings. However, three of the pillars supporting the proposed framework are very known there: Feature Subset Selection, Dimensionality Reduction and Data Visualization. Thus, in this section a short overview about theses points is presented.

Feature Subset Selection (FSS) problem [1] deals with the search of the best subset of attributes to train a classifier. This is a very important issue in several areas of knowledge discovery such as machine learning, optimization, pattern recognition and statistics. The goal behind FSS is the appropriate selection of a relevant subset of features upon which to focus the attention of a classification algorithm, while ignoring the rest. The FSS problem is based on the fact that the inclusion of more attributes in a training dataset does not necessarily improve the performance of the model. Two different kinds of variables can be distinguished: irrelevant (variable has no relation with the target of the classifier) and redundant (variables whose information can be deduced from other variables) features.

The literature describes several approaches to tackle this problem. To achieve the best possible performance with a particular learning algorithm on a particular training set, a FSS method should consider how the algorithm and the training set interact. Thus, there are two alternatives to consider this interaction: Filter [2] (analytical and statistical information among the features to evaluate each available feature) and Wrapper methods [3] (they use the induction algorithm itself to evaluate the performance of each candidate feature selection).

Wrapper methods often achieve better feature selections but the computational cost is higher. There are two main aspects that influence deeply on the computational cost of these techniques: (i) the optimization algorithm could be more or less exhaustive. For example, forward selection, backward elimination, and their stepwise variants can be viewed as simple hill-climbing techniques in the space of feature subsets; (ii) the robustness of the validation method (for instance LOOCV, Bootstrap, ...) applied to evaluate the quality of the results obtained by each candidate selection. It includes the measure to use, but also the validation schema.

An accurate FSS technique based on wrapper approaches that combines both a powerful search method and a robust validation approach is still a challenge, particularly in high-dimensional datasets. An appropriate alternative is to use a hybrid approach [4,5].

The most common one is the use of a filter to reduce the number of features (features are ranked based on their representativeness and the worst ones are removed), and a wrapper to perform the final selection. This represents a balance between the number of features to make the wrapper technique reasonable in computational time and the number of features included in the optimal subset selection.

As regards Data Visualization (DV), and specifically Multidimensional (unknown relations between attributes) Multivariate Data Visualization (MMDV), there are four broad categories [6] according to the approaches taken to generate the resulting visualizations. The first, *Geometric projection*, includes techniques that aim to find informative projections and transformations of multidimensional datasets [7] such as the Scatterplot Matrix [8], the Prosection Matrix [9], Parallel Coordinates [10] and Star Coordinates [11]. The second category groups the *Pixel-oriented* techniques [7] that represent a feature value by a pixel based on a color scale. This group includes the Space Filling Curve [12], the Recursive Pattern [13] and Spiral and Axes Techniques [14], among others. The techniques of the third category, *Hierarchical techniques*, subdivide the data space and present sub-spaces in a hierarchical way [7], for example, the Hierarchical Axis [15] and Dimensional Stacking [16] methods. The last category, *Iconography*, represents icon-based techniques that map the multidimensional data to different icons, or glyphs [17]. Some of them are Chernoff Faces [18] and Star Glyph [19].

Another way of visualizing multidimensional and multivariate data is by carrying out a Dimensionality Reduction (DR) process, which is one of the usual operations in Data Analysis (DA) [20]. Historically, the main reasons for reducing the dimensionality of the data is to remove possible noise or redundancy in the data, and reducing the computational load in further processing. One of the fields in which DR techniques for DV are currently very useful, is the scientific interactive visualization field, or Visual Analytics (VA). For DV, one of the main applications of DR is to map a set of observations into a 2 or 3 dimensional space that preserves the intrinsic geometric structure of the data as much as possible [21]. More related work about DR is presented in [22].

3 MedVir

The *MedVir* framework has been devised to abstract the clinicians the slow and tedious task of extracting conclusions about patients, treatments and rehabilitation when they work with multidimensional multivariate data analysis. The idea is that the expert only has to select the data to work with, and MedVir carries out a pipeline containing the most important steps of the KDD (Knowledge Discovery in Databases) process. As a result, data can be easily visualized in a virtual environment allowing a complete interaction, in order to get more conclusions about the interests of the clinicians.

MedVir comprises the following stages, as illustrated in Figure 1: i) *data preprocessing*, in which a set of data transformations and formatting are carried out so that the data can be properly treated by the following steps; ii) *selection of a reduced number of attributes* that best describe the original nature of the dataset. This step is carried out by using an extensive and intensive FSS process, in which five filter methods, four wrapper methods and four classification algorithms are used to obtain the models

that perform better in supervised learning tasks; iii) *dimensionality data reduction* up to 2 or 3 dimensions to correctly represent the data on the display, with a minimum loss of quality; iv) *visualization of the data* facilitating a quick data interpretation.

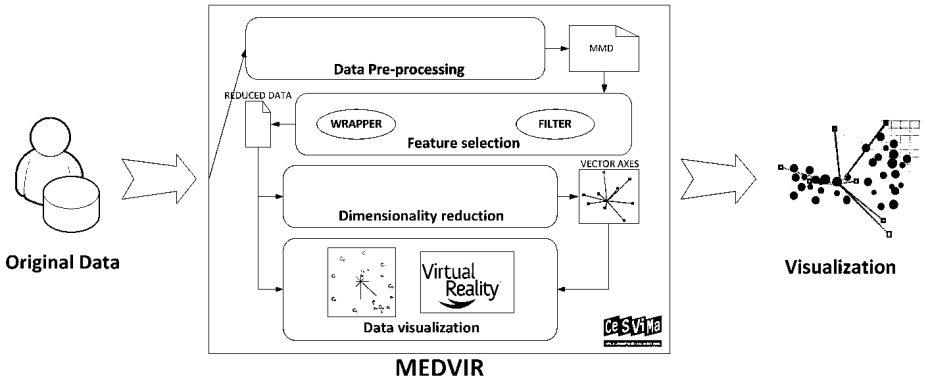

Fig. 1. The MedVir framework

Data Pre-processing. Real data often have a lot of redundancy, as well as incorrect or missing values, depending on different factors. Thus, it is usually necessary to perform some techniques in order to clean up and prepare the data. The algorithms included in this stage are replicated features handling, missing value handling and imputing missing values with *KNNImpute* algorithm [23]. Real data often have a lot of redundancy, as well as incorrect or missing values, depending on different factors. Thus, it is usually necessary to perform some techniques in order to clean up and prepare the data. The algorithms included in this stage are replicated features handling, missing value handling and imputing missing values with *KNNImpute* algorithm [23].

Feature Subset Selection. The second stage consists of a FSS process, which is responsible for selecting a reduced subset of attributes, from a very large number of initial attributes. The aim is to obtain a reduced dataset that retains or improves efficiency in many different Data Mining tasks. Thus, the main advantage of this stage is that the number of data attributes are strongly reduced from tens of thousands to a few dozens of attributes, thus reducing the computational cost and retaining or even improving their accuracy in different tasks, such as supervised or unsupervised classification. It is worth mentioning that the study presented here is limited to supervised classification tasks.

This step consists of two sub-stages: *filter* and *wrapper*. To implement the filter approach, five filter methods were used (Information gain, ReliefF, Symmetrical Uncertainty, Gain ratio and Chi squared) [24,25], and each one of these is executed *P times*, that is, for the different numbers of attributes to be filtered (eg, 500, 1000, 2000, ...). Once the filtered dataset is obtained, a wrapper process is carried out, using four search methods (Greedy, Best first, Genetic and Linear forward selection (LFS)) [24,25] and four classification algorithms (C4.5, SVM, Bayes Net and K-NN) [24,25] to obtain a reduced dataset containing, most of the cases, a few dozens of attributes. The combined

use of wrapper and filter methods generate $80P$ different models and those that produces the best values, in terms of accuracy, are selected. To validate the results of each model, the *0.632 Bootstrap* [26] validation method has been used. Note that P can be set according to the number of attributes contained in the original data (e.g., if the dataset has 5000 attributes, P executions could be 6: 500, 1000, 2000, 3000, 4000 and 4500).

Dimensionality Reduction. The optimal dataset obtained in the previous stage can still not be directly visualized in two or three dimensions, since in many cases these data are supposed to have more than 3 attributes. We say optimal because, at this point, a dataset with a minimum number of attributes has been obtained, which always preserves or even improves (never worsens) efficacy when carrying out different tasks. Therefore, the third stage is responsible for obtaining a set of vector axes (generated by a particular DR algorithm) to be used in the next stage of MedVir's pipeline, so that the reduced data are transformed to be visualized properly in 2 or 3 dimensions.

Different DR algorithms can be, indeed, used at this stage. For example, for clustering tasks, one might be interested in using PCA, since due to its great ability to obtain the directions of maximum variance of data, it produces minimum loss of quality of data [22], thus making more reliable the visualization of the real structure of data. Instead, LDA could be useful for supervised tasks, because even if the effectiveness in the preservation of the original geometry data is drastically reduced [22], the spatial directions of maximum discrimination between classes are easily obtained. This will facilitate the separation of different classes when the data are displayed. Therefore, depending on the used DR algorithm, a set of vectors are generated (as many vectors as attributes has the reduced dataset prior to this stage) to be used in the last stage of MedVir.

Data Visualization. The last stage generates the final visualization of the reduced data. To do so, the *star coordinates* (SC) algorithm is used [27]. SC algorithm works as follows: first, each attribute is represented as a vector radiating from the center of a circle to its circumference. Then the coordinate axes are arranged onto a flat (2-dimensional) surface forming equidistant angles between the axes. The mapping of an D-dimensional point to a 2-dimensional Cartesian coordinate is computed by means of the sum of all unit vectors on each coordinate, multiplied by the data value for that coordinate. In this paper, the input to the SC algorithm comprises two different elements: the *reduced data* and the set of *vector axes* generated in the previous stage. Thus, final visualization will be adjusted to the DR algoritm's requirements.

The MedVir's *visualization* and *interaction* comprise, among many others, the carrying out of the following tasks: 1) if two points of different class (color) are very close or even overlapped in the visualization. This could strongly suggest that the expert might have made a mistake when originally labelling those instances. 2) if an attribute is selected, all points will be resized based on that attribute's value. This could represent the *importance* or *influence* of that attribute on a particular class. 3) if one or more attributes are selected and their lengths are modified, we would be giving them more or less weight on the representation, so all the instances will be reorganized based on those new weights. For example, if we give a greater weight to an attribute and a point with class A approaches another point with class B, this could suggest that a higher value

of that attribute will mean a change in instance status from class A to B. 4) selected attributes can be *removed* to reduce their influence on the data representation. 5) if the instances are patients, their clinical information can be visualized *quickly* and *easily*. 6) a different visual *dispersion* among members of the same class and other classes may suggest different levels of cohesion between different instances. 7) display can be adjusted to achieve a comfortable interface, and the 2D and 3D visualization is represented by different colors, sizes, transparencies and shapes. Furthermore, navigation is *simple* and *intuitive*.

4 MedVir Applied to TBI

MedVir was applied to a real world case (figure 2), that is a Traumatic Brain Injury (TBI) rehabilitation prediction [28]. The study was performed by 12 control subjects and 14 patients with brain injury. All patients have completed a neurorehabilitation program, which was adapted specifically to each individual's requirements. This program was conducted in individual sessions attempting to offer an intensive neuropsychological-based rehabilitation, provided in 1h sessions for 3-4 days a week. In some cases, cognitive intervention was coupled with other types of neurorehabilitation therapies according to the patient's profile.

Patients had MEG recordings before and after the neuropsychological rehabilitation program. In this study control subjects were measured once, assuming that brain networks do not change in their structure in less than one year, as demonstrated previously in young.

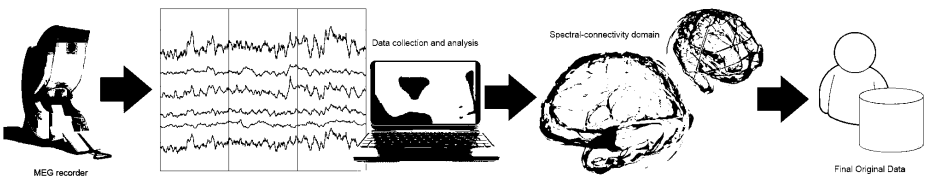

Fig. 2. MEG data obtaining process

The magnetic fields were recorded using a 148-channel whole-head magnetometer confined in 40 magnetically shielded room. MEG data were submitted to an interactive environmental noise reduction procedure. Fields were measured during a no task eyes-open condition. Time-segments containing eye movements or blinks or other myogenic or mechanical artefacts were rejected and time windows not containing artefacts were visually selected by experienced investigators, up to a segment length of 12s. By using a wavelet transformation [29], we perform a time-frequency analysis of rhythmic components in a MEG signal, and hence estimate the wavelet coherence for a pair of signals, a normalized measure of association between two time series [30]. Finally, MEG data were digitalized and transformed into a simple dataset of 26 instances x 10878 attributes, where each instance is a patient and each attribute is the relationship between each pair of channels.

4.1 Experiments

Experiments on the aforementioned dataset consists of a *FSS* process and *visualization* of the obtained data.

The first stage, *FSS*, is responsible for selecting, from among the 10,878 original attributes, a set of reduced data which improve accuracy when classifying new patients, compared to the original data. To classify new patients, a specific dataset consisting of 14 new instances is used. The FSS process consists of two parts: the first one uses filter methods, and the second one uses wrapper methods on the previous filtered attributes.

In total, 480 different models (5 *filter methods* x 6 *number of attributes to be filtered* x 4 *search methods* x 4 *classification algorithms*) have been obtained over the two parts, of which those who obtained the best accuracy were selected. The aim was to apply those models to the data to eventually visualize them. Note that the P value, described in Section 3 has been set to 6, since each filter method is carried out on the 500, 1000, 2000, 3000, 4000 and 5000 best attributes. The implementation of these models has been carried out in parallel and using the Magerit supercomputer, thus 480 nodes of the supercomputer have been used simultaneously to obtain the results. At the end of the process, a ranking of the 480 models was obtained, sorted by time spent to carry out the experiments (see Figure 3). Furthermore, the results have been validated using 0.632 Bootstrap method, as indicated in Section 3.

	Model	% Accuracy (Original)	Accuracy (Filtered)	Accuracy (Wrappered)	N° of attributes	Time (s)
↑	TBI_Relieff_500_Genetic_KNN	64.546	63.996	71.168	10	578.041
480	TBI_SymmetricalUncert_5000_Genetic_SVM	67.893	63.411	72.243	65	6126.37
↓	...	...	...	...	...	...
	...	...	...	...	...	...

Fig. 3. An example of the ranking of models obtained after the FSS stage

The criterion to select the best models is based on the highest values of accuracy achieved after the carrying out of the wrapper methods (fourth column from left). So, the models that have obtained the best accuracy are:

- **TBI_Relieff_500_Genetic_KNN (71.16%).** The first model has used the *relieff* filter to obtain the best 500 attributes. After this, a *genetic* algorithm carried out an extensive search to select a subset of the 10 best attributes that best discriminate between the original classes, when classifying the instances by using the *K-NN* classification algorithm.
- **TBI_SymmetricalUncert_5000_Genetic_SVM (72.24 %).** The second model has used the *symmetrical uncert* filter to rank the best 5000 attributes. Then, a *genetic* algorithm has selected a subset of the 65 best attributes that best discriminate between the original classes, when using the *SVM* classification algorithm.

Therefore, once the two reduced datasets were obtained, the classification of new patients was carried out. The results of the classification task are shown in Figure 4 (0 represents control subjects and 1 represents TBI patients). Except for patients 3 and 4 contained in the test dataset, there is a clear unanimity between the classification carried out by both models.

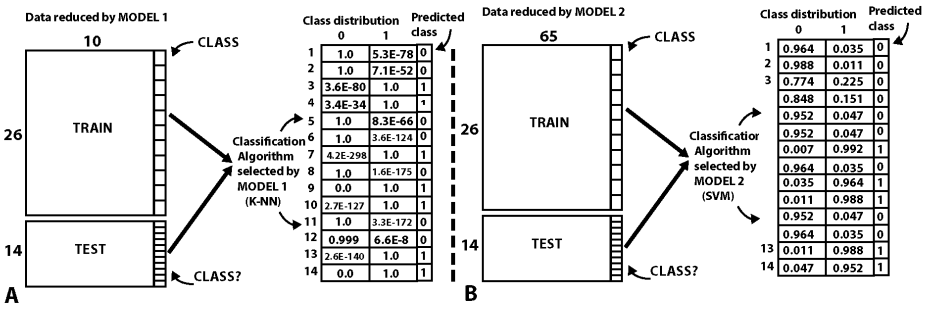

Fig. 4. Two models to classify the new patients

In the last step, *visualization*, MedVir represents the two datasets (figure 5). There, the blue color represents control subjects, whiles red means TBI patients and new classified patients are represented in magenta. The dotted line indicates the linear decision boundary in classification tasks. And it is at this time, when experts analyze the resulting work in order to extract the maximum possible information and draw relevant conclusions.

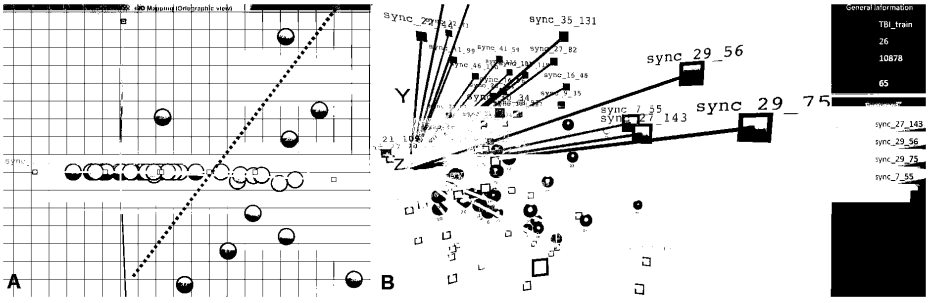

Fig. 5. Visualization in MedVir

5 Conclusions

MedVir can visualize multidimensional and multivariate medical data in 3D, so this allows conclusions to be obtained in a more simple, quick and intuitive way. Furthermore, the use of MedVir allows the clinicians to interact with the data they collect daily.

Specifically for this study, MedVir allows to effectively segment between control subjects and TBI patients with a 72% of accuracy. This is not a very high value, but it must be, indeed, taked into account that the validation mechanism used (Bootstrap) is certainly pessimistic due to the small number of instances in the data, so this strongly penalizes the final accuracy. In this paper, MedVir has been presented as a quick and easy tool to classify and visualize new subjects included in the TBI study. Visualization and interaction with the data can provide extra useful information to discern between uncertain class patients, after obtaining the results of a classification. In addition, MedVir could even be used to estimate if a TBI patient is in process of rehabilitation or not, so clinicians could be able to change the treatment or stop it.

However, there will be further research behind this work. In terms of data analysis, regression models and neuropsychological tests are going to be included to estimate the exact situation of a TBI patient in the recovery process, and how much treatment time he will need to fully rehabilitate. Another interesting future research point is the DR of data based on clustering of MEG sensors (creation of brain regions based on sensors locations and their relationships). In terms of visualization, we want to improve the user's interaction, using IO devices such as Leap Motion (MedVir controlled by gestual movements) and voice recognition (by means of expert orders). Furthermore, the interaction with the data visualization should be further studied by carrying out usability tests to test its reliability.

Concluding, MedVir, as a analysis tool, has successfully served for its purpose, allowing to know the status of rehabilitation of the TBI patients in an easy way. Of course, this tool could be applied to another interesting field, in which the number of attributes are too high that makes impossible a direct data analysis.

References

1. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271 (1997)
2. Jeffery, I.B., Higgins, D.G., Culhane, A.C.: Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 7, 359+ (2006)
3. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
4. Ni, B., Liu, J.: A hybrid filter/wrapper gene selection method for microarray classification. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2537–2542. IEEE (2004)
5. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial Intelligence in Medicine 31, 91–103 (2004)
6. Keim, D.A.: Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics 8, 1–8 (2002)
7. Keim, D.A., Kriegel, H.P.: Visualization Techniques for Mining Large Databases: A Comparison. Transactions on Knowledge and Data Engineering, Special Issue on Data Mining 8, 923–938 (1996)
8. Hartigan, J.: Printer graphics for clustering. Journal of Statistical Computation and Simulation 4, 187–213 (1975)
9. Furnas, G.W., Buja, A.: Prosection Views: Dimensional Inference through Sections and Projections. Journal of Computational and Graphical Statistics 3, 323–385 (1994)
10. Inselberg, A.: Multidimensional Detective. In: Proceedings of the 1997 IEEE Symposium on Information Visualization, INFOVIS 1997, pp. 100–107. IEEE Computer Society, Washington, DC (1997)
11. Beddow, J.: Shape Coding of Multidimensional Data on a Microcomputer Display. In: IEEE Visualization, pp. 238–246 (1990)
12. Peano, G.: Sur une courbe, qui remplit toute une aire plane. Mathematische Annalen 36, 157–160 (1890)
13. Keim, D.A., Ankerst, M., Kriegel, H.P.: Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data. In: Proceedings of the 6th Conference on Visualization 1995, VIS 1995, pp. 279–286. IEEE Computer Society, Washington, DC (1995)

14. Keim, D.A., Krigel, H.P.: VisDB: Database Exploration Using Multidimensional Visualization. IEEE Comput. Graph. Appl. 14, 40–49 (1994)
15. Mihalisin, T., Gawlinski, E., Timlin, J., Schwegler, J.: Visualizing a Scalar Field on an N-dimensional Lattice. In: Proceedings of the 1st Conference on Visualization 1990, VIS 1990, pp. 255–262. IEEE Computer Society Press, Los Alamitos (1990)
16. LeBlanc, J., Ward, M.O., Wittels, N.: Exploring N-dimensional Databases. In: Proceedings of the 1st Conference on Visualization 1990, VIS 1990, pp. 230–237. IEEE Computer Society Press, Los Alamitos (1990)
17. de Oliveira, M.C.F., Levkowitz, H.: From Visual Data Exploration to Visual Data Mining: A Survey. IEEE Trans. Vis. Comput. Graph. 9, 378–394 (2003)
18. Chernoff, H.: The Use of Faces to Represent Points in K-Dimensional Space Graphically. Journal of the American Statistical Association 68, 361–368 (1973)
19. Chambers, J., Cleveland, W., Kleiner, B., Tukey, P.: Graphical Methods for Data Analysis. The Wadsworth Statistics/Probability Series. Duxury, Boston (1983)
20. Lee, J.A., Verleysen, M.: Nonlinear dimensionality reduction. Springer, New York (2007)
21. Wang, J.: Geometric Structure of High-dimensional Data and Dimensionality Reduction. Higher Education Press (2012)
22. Gracia, A., González, S., Robles, V., Menasalvas, E.: A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. Information Sciences (2014)
23. Speed, T.: Statistical analysis of gene expression microarray data. CRC Press (2004)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11, 10–18 (2009)
25. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers Inc., San Francisco (2005)
26. Efron, B., Tibshirani, R.: Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association 92, 548–560 (1997)
27. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: KDD 2001: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 107–116. ACM, New York (2001)
28. Castellanos, N.P., Paul, N., Ordonez, V.E., Demuynck, O., Bajo, R., Campo, P., Bilbao, A., Ortiz, T., del Pozo, F., Maestu, F.: Reorganization of functional connectivity as a correlate of cognitive recovery in acquired brain injury. Brain Journal 133, 2365–2381 (2010)
29. Mallat, S.: A Wavelet Tour of Signal Processing, The Sparse Way, 3rd edn. Academic Press (2008)
30. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc. 79, 61–78 (1998)