Comparative Diagnostic Accuracy of Linear and Nonlinear Feature Extraction Methods in a Neuro-oncology Problem

Raúl Cruz-Barbosa[1], David Bautista-Villavicencio[1], and Alfredo Vellido[2]

[1] Universidad Tecnológica de la Mixteca, 69000, Huajuapan, Oaxaca, México
{rcruz,dbautista}@mixteco.utm.mx
[2] Universitat Politècnica de Catalunya, 08034, Barcelona, Spain
avellido@lsi.upc.edu

Abstract. The diagnostic classification of human brain tumours on the basis of magnetic resonance spectra is a non-trivial problem in which dimensionality reduction is almost mandatory. This may take the form of feature selection or feature extraction. In feature extraction using manifold learning models, multivariate data are described through a low-dimensional manifold embedded in data space. Similarities between points along this manifold are best expressed as geodesic distances or their approximations. These approximations can be computationally intensive, and several alternative software implementations have been recently compared in terms of computation times. The current brief paper extends this research to investigate the comparative ability of dimensionality-reduced data descriptions to accurately classify several types of human brain tumours. The results suggest that the way in which the underlying data manifold is constructed in nonlinear dimensionality reduction methods strongly influences the classification results.

1 Introduction

The diagnostic classification of human brain tumours on the basis of single-voxel proton magnetic resonance spectroscopy (SV-^{1}H-MRS) information is a non-trivial problem in which dimensionality reduction (DR) is almost mandatory [1]. DR strategies usually take the form of feature selection or feature extraction [2]. In feature extraction using manifold learning models [3], multivariate data are described through a low-dimensional manifold embedded in data space.

Although the Euclidean metric is often used in this setting, similarities between points along the underlying manifold have been shown to be best expressed as geodesic distances or their approximations [4–7]. This is specially important if working with high-dimensional data of unknown intrinsic geometry. Such approximations of the geodesic distances along the manifold can be computationally intensive, and several alternative software implementations of manifold learning models have been recently put forward and compared in terms of their computation times, using several standard and non-standard data sets as benchmarks [8].

J.-F. Martínez-Trinidad et al. (Eds.): MCPR 2011, LNCS 6718, pp. 34–41, 2011.

Some of the proposed computational time-saving strategies showed great promise in the sense that they were fast while not compromising the amount of data variance explained. They were bundled in a software module that was inserted in a nonlinear dimensionality reduction (NLDR) method, namely ISOMAP [4]. The current brief paper moves this research one step forward to investigate the comparative ability of these dimensionality-reduced data descriptions to accurately classify several types of human brain tumours on the basis of SV-^{1}H-MRS information. The performance of the most computationally-effective method is compared to that of two alternative ISOMAP implementations, and to the well-known Principal Component Analysis (PCA) linear technique. Classification is carried out using a simple linear method, namely Linear Discriminant Analysis (LDA), which has previously been successfully applied to this type of data. The results suggest that the way in which the data manifold is constructed in ISOMAP compromises the achieved classification accuracy, although one of the alternative ISOMAP implementations provides, in some of the experiments, comparable accuracy results to those of PCA with fewer features.

2 Methods

2.1 Optimizing the Computation of Geodesic Distances

The explicit calculation of geodesic distances can be computationally impractical. This metric, though, can be approximated by graph distances [9], so that instead of finding the minimum arc-length between two data points lying on a manifold, we would set to find the shortest path between them, where such path is built by connecting the closest successive data points. In this paper, this is done using the K-rule, which allows connecting the K-nearest neighbours. A weighted graph is then constructed by using the data and the set of allowed connections. The data are the vertices, the allowed connections are the edges, and the edge labels are the Euclidean distances between the corresponding vertices. If the resulting graph is disconnected, some edges are added using a minimum spanning tree procedure in order to connect it. Finally, the distance matrix of the weighted undirected graph is obtained by repeatedly applying Dijkstra's algorithm [10], which computes the shortest path between all data samples. This process is illustrated in Fig. 1.

There are different implementation choices for some of the stages involved in the geodesic distance computation (see Fig. 1). Two alternatives for graph representation are the *adjacency matrix* and the *adjacency list*. The former consists in a n by n matrix structure, where n is the number of vertices in the graph. If there is an edge from a vertex i to a vertex j, then the element a_{ij} is 1, otherwise it is 0. This kind of structure provides faster access for some applications but can consume huge amounts of memory. The latter considers that each vertex has a list of which vertices it is adjacent to. This structure is often preferred for sparse graphs as it has smaller memory requirements. Three options for the shortest path algorithm are the standard Dijkstra, Dijkstra using a Fibonacci heap (F-heap) and Floyd-Warshall. For some applications where the obtained graph is

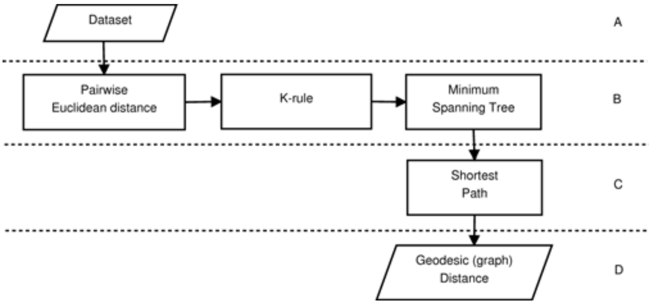

Fig. 1. Graph distance procedure scheme. Stage (A) represents the input data. Stage (B) is for building the weighted, undirected, connected graph. Stage (C) is for computing the geodesic (graph) distance, which is returned in Stage (D).

a sparse graph, Dijkstra's algorithm can save memory resources by storing the graph in the form of adjacency list and using a F-heap [11] as a priority queue, reducing the time complexity of the algorithm.

Previous research [8] assessed which combination of graph representation and shortest path algorithm produced the best time performance for computing the geodesic distance for datasets with increasing numbers of items. Alternative C++ and Matlab® implementations were also tested.

A combination of adjacency matrix for graph representation and basic Dijkstra as the choice for shortest path algorithm outperformed the other combinations in most settings, due to the faster access to elements in an adjacency matrix, especially for small data sets (as the ones analyzed in this study). The time performance of the C++ implementation of the geodesic distance computation using a matrix representation and the basic Dijkstra algorithm clearly outperformed its Matlab® counterpart. All experiments were performed using a dual-processor 2.3 Ghz BE-2400 desk PC with 2.7Gb RAM.

2.2 Dimensionality Reduction and Classification: ISOMAP and LDA

One way in which DR through feature extraction methods can be categorized is as linear or nonlinear ones. One of the best-known linear methods (and also one that is widely applied to biomedical problems) is PCA. The main aim of PCA is reducing the data dimensionality through an orthogonal transformation, while retaining as much as possible of the data variance along the main extracted dimensions (components).

A recent NLDR method that is increasingly gaining in popularity is ISOMAP [4]. This method is a variant of multi-dimensional scaling, which aims to embed high dimensional data points onto a low dimensional space by preserving interpoint distances as closely as possible. In this method, the geodesic (graph) metric is proposed to compute distances along the manifold instead of the Euclidean one.

The extracted features are then amenable to classification analysis. A basic but efficient classifier (and, again, one commonly used in the analysis of medical data) is LDA. It aims to find a linear combination of features that optimally characterizes or separates different data classes.

3 Materials: SV-^{1}H-MRS Brain Tumour Database

The experiments in this study concern MRS data acquired at different echo times (short time of echo -STE- and long time of echo -LTE-), as well as with a combination of both by straightforward concatenation. Data belong to a multi-center, international database [12], and consist of: (1) 217 STE spectra, including 58 meningiomas (mm), 86 glioblastomas (gl), 38 metastases (me), 22 astrocytomas grade II (a2), 6 oligoastrocytomas grade II (oa), and 7 oligodendrogliomas grade II (od); (2) 195 LTE spectra, including 55 mm, 78 gl, 31 me, 20 a2, 6 oa, and 5 od. (3) 195 items built by combination (through direct concatenation) of the STE and LTE spectra for the same patients. Only the clinically relevant regions of the spectra were analyzed. They consist of 195 frequency intensity values (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector), starting at 4.25 ppm. These frequencies become the observed data features.

The classification experiments involved grouping these tumour types into three classes , namely: G1: low-grade gliomas (a2, oa and od); G2: high-grade malignant tumours (me and gl); and G3: meningiomas (mm). This type of grouping is justified [1, 13] by the well-known difficulty in distinguishing between metastases and glioblastomas, due to their similar spectral pattern.

4 Results and Discussion

The goal of the experiments reported in this section is twofold. Firstly, we aim to show how the way the data manifold is constructed by the ISOMAP model variants affects the classification accuracy. Secondly, we aim to assess the classification results in terms of the number of extracted features and the corresponding accuracy. For all experiments, the K parameter for the K-rule is set to 10 in order to get a connected graph after this rule is applied. The three ISOMAP variants investigated are: a computationally-optimized one implemented in C++ (namely ISOMAP gMod), obtained by embedding the geodesic distance calculation software module described in [8]; and two variants of Tenenbaum's [4] ISOMAP implementation (standard and landmark).

The average classification accuracy results are validated by 10-fold cross-validation. The LDA classification results for G1 vs G2 vs G3 using STE, LTE and STE+LTE spectra are summarily reported in Figs. 2, 3 and 4. The accuracies achieved by the different ISOMAP implementations neatly differ. These differences are the result of the different manners in which the underlying manifold graph representations are obtained by the corresponding variants of the algorithm. ISOMAP standard takes the largest component when the resulting

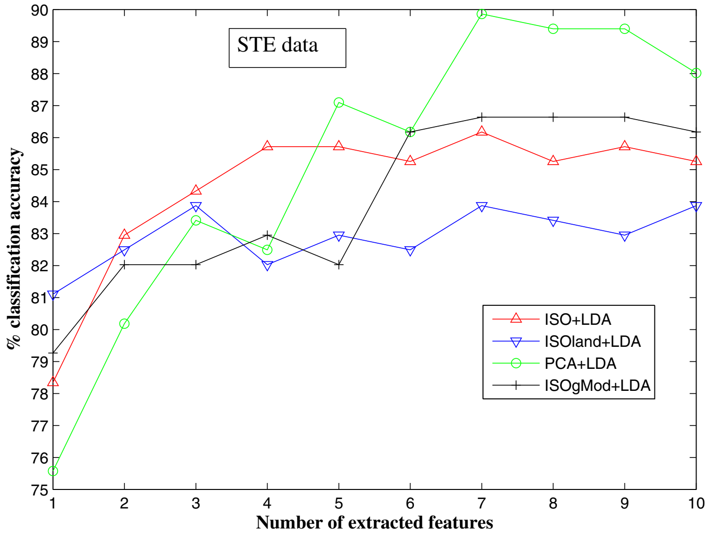

Fig. 2. LDA classification results for G1 vs G2 vs G3 using features extracted from STE spectra. ISOMAP variants: standard (ISO), landmark(ISOland) and with the computationally-optimized module (ISOgMod).

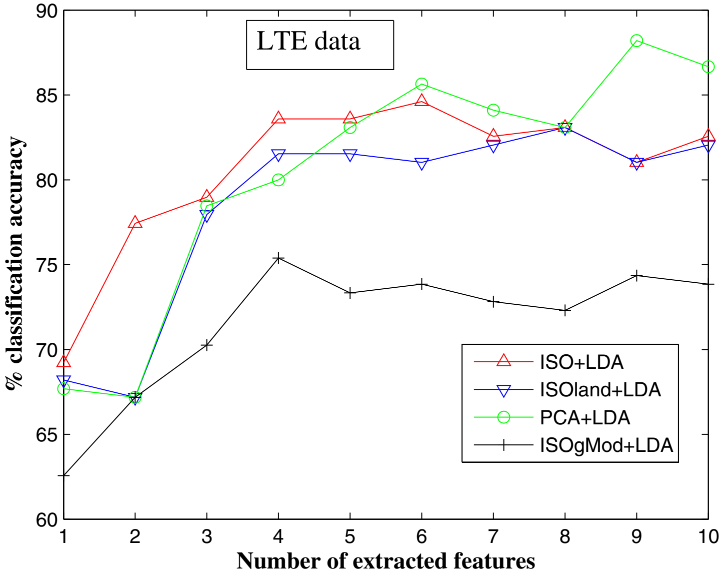

Fig. 3. LDA classification results for G1 vs G2 vs G3 using features extracted from LTE spectra. Legend as in Fig. 2.

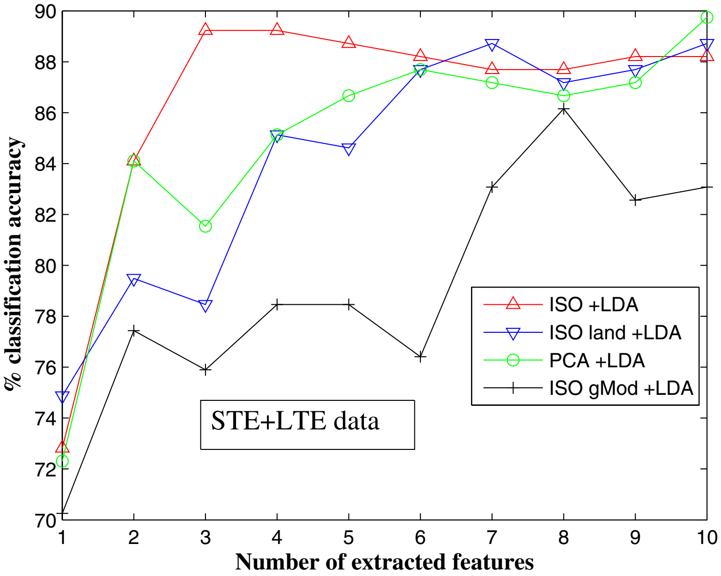

Fig. 4. LDA classification results for G1 vs G2 vs G3 using features extracted from STE+LTE spectra. Legend as in Fig. 2.

graph, created by the K-rule, is disconnected. ISOMAP gMod, instead, always builds a connected graph by using a modified minimum spanning-tree procedure, realized by connecting the closest data points between the disconnected components. Finally, ISOMAP landmark randomly selects l landmark points from the original data and the graph is constructed using only those points. Thus, the resulting graph-based distance matrix is strongly dependent on the way the graph is constructed. Since ISOMAP uses the graph distance matrix as input for a multidimensional scaling method which computes the coordinates of the data points in the low dimensional projection space, the extracted features differ notably, hence different classification accuracy results are obtained.

The results reported in Figs. 2 and 3 indicate that, for small numbers of extracted features, the ISOMAP implementations do not provide significantly better LDA accuracies than PCA. For 5 or more features, PCA clearly outperforms ISOMAP to reach just under 90% average accuracy. ISOMAP variants only outperform PCA when data combining the two times of echo are used (as seen in Fig. 4). An extremely parsimonious data representation consisting of just 3 features is enough to obtain an average accuracy just below 90%. Overall, these classification accuracy results are consistent with those reported in literature. For example, in [14], the use of STE+LTE spectra with a PCA+LDA setting achieved a classification accuracy of around 90%, but using a minimum of 8 principal components. The classification results for ISOMAP with gMod implementation deteriorate significantly whenever LTE spectra are used.

5 Conclusion

Brain tumours show a relatively low prevalence in the general population, but their diagnosis and prognosis are challenging and sensitive medical issues. Machine learning and computational intelligence methods can assist medical doctors and expert radiologists in these tasks [15, 16]. The classification of human brain tumours on the basis of high-dimensional SV-^{1}H-MRS makes the use of DR strategies advisable. Manifold learning techniques using geodesic distances have previously shown promise in this DR task, and there have been efforts to optimize their intensive use of computational resources.

In this paper, we have focused in the ISOMAP NLDR method. Previous research [8] provided evidence that a specific implementation of a geodesic distance computation module in C++ language, as part of the ISOMAP implementation, had an extremely fast performance. In this study, we have carried out preliminary experiments to gauge the ability of the data reduction obtained with the most computationally-effective implementation to provide accurate diagnostic classification for several common brain tumour pathologies on the basis of MRS data. The results reported in this paper show that this most computationally-effective implementation does not perform well in many of the experimental settings. This is evidence that the way in which the data manifold is constructed in NLDR manifold learning methods may compromise the subsequent classification accuracy. The standard ISOMAP implementation, though, is still capable of achieving maximum accuracy in the brain tumour classification problem with far less features than PCA, if a combination of data at different times of echo is used. Further research should include more types of brain tumours, as well as a wider palette of NLDR manifold learning techniques. The interpretability of the features extracted by NLDR, from a medical point of view, should also be investigated.

Acknowledgments. Partial funding for this research was provided by the Mexican SEP PROMEP/103.5/10/5058 and the Spanish MICINN TIN2009-13895-C02-01 research projects. Authors gratefully acknowledge the former INTERPRET European project partners. Data providers: Dr. C. Majós (IDI), Dr. À. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Prof. L. Stefanczyk and Dr J. Fortuniak (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. M. Julià-Sapé, Dr. A.P. Candiota, Dr. I. Olier, Ms. T. Delgado, Ms. J. Martín and Mr. A. Pérez (all from GABRMN-UAB). GABRMN coordinator: Prof. C. Arús.

References

1. Vellido, A., Romero, E., González-Navarro, F., Belanche-Muñoz, L., Julià-Sapé, M., Arús, C.: Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database. Neurocomputing 72(13-15), 3085–3097 (2009)
2. González-Navarro, F., Belanche-Muñoz, L., Romero, E., Vellido, A., Julià-Sapé, M., Arús, C.: Feature and model selection with discriminatory visualization for diagnostic classification of brain tumours. Neurocomputing 73(4-6), 622–632 (2010)

3. Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, Heidelberg (2007)
4. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
5. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
6. Roweis, S.T., Lawrence, K.S.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
7. Cruz-Barbosa, R., Vellido, A.: Semi-supervised geodesic generative topographic mapping. Pattern Recognition Letters 31(3), 202–209 (2010)
8. Bautista-Villavicencio, D., Cruz-Barbosa, R.: On geodesic distance computation: An experimental study. Advances in Computer Science and Applications, Research in Computing Science 53, 115–124 (2011)
9. Bernstein, M., de Silva, V., Langford, J.C., Tenenbaum, J.B.: Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, CA, U.S.A (2000)
10. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959)
11. Fredman, M.L., Tarjan, R.E.: Fibonacci heaps and their uses in improved network optimization algorithms. J. ACM 34(3), 596–615 (1987)
12. Julià-Sapé, M., et al.: A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. Magn. Reson. Mater. Phys. MAGMA 19, 22–33 (2006)
13. Tate, A.R., Majós, C., Moreno, A., Howe, F.A., Griffiths, J.R., Arús, C.: Automated classification of short echo time in In Vivo ^{1}H brain tumor spectra: a multicenter study. Magnetic Resonance in Medicine 49, 29–36 (2003)
14. García-Gómez, J.M., Tortajada, S., Vidal, C., Julià-Sapé, M., Luts, J., Moreno-Torres, A., Van-Huffel, S., Arús, C., Robles, M.: The effect of combining two echo times in automatic brain tumor classification by MRS. NMR in Biomedicine 21(10), 1112–1125 (2008)
15. Lisboa, P.J.G., Vellido, A., Tagliaferri, R., Napolitano, F., Ceccarelli, M., Martin-Guerrero, J.D., Biganzoli, E.: Data mining in cancer research. IEEE Computational Intelligence Magazine 5(1), 14–18 (2010)
16. Vellido, A., Lisboa, P.J.G.: Neural networks and other machine learning methods in cancer research. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 964–971. Springer, Heidelberg (2007)