X-SDR: An Extensible Experimentation Suite for Dimensionality Reduction

Panagis Magdalinos, Anastasios Kapernekas, Alexandros Mpiratsis, Michalis Vazirgiannis

Athens University of Economics and Business Athens, Greece
pmagdal@aueb.gr,kapernekas@aueb.gr,abiratsis@gmail.com,mvazirg@aueb.gr

Abstract. Due to the vast amount and pace of high-dimensional data production, dimensionality reduction emerges as an important requirement in many application areas. In this paper, we introduce X-SDR, a prototype designed specifically for the deployment and assessment of dimensionality reduction techniques. X-SDR is an integrated environment for dimensionality reduction and knowledge discovery that can be effectively used in the data mining process. In the current version, it supports communication with different database management systems and integrates a wealth of dimensionality reduction algorithms both distributed and centralized. Additionally, it interacts with Weka thus enabling the exploitation of the data mining algorithms therein. Finally, X-SDR provides an API that enables the integration and evaluation of any dimensionality reduction algorithm.

Keywords: dimensionality reduction, data mining, knowledge discovery

1 Introduction

Data pre-processing is a crucial step of data mining that enables the abduction of irrelevant values from a dataset. An important aspect of data pre-processing is dimensionality reduction (DR). DR methods address challenges that rise from the large number of variables describing each observation. In high dimensional spaces typical knowledge discovery tasks, such as clustering or classification become ineffective [1]. DR algorithms apply transformations on the original dataset and embed it from the original space R^n to a new, low dimensional space R^k ($k << n$). The objective of the methodology is to retain the distances between points or other statistical properties (i.e. variance) in the new space. Recently, due to the advent of large distributed applications, distributed dimensionality reduction (DDR) has also emerged as a decentralized pre-processing step.

Despite the large number of centralized ([4]) and distributed ([8] and references therein) approaches we lack a software tool that integrates them towards experimenting with them in a user friendly manner. The latter, has inspired the design and implementation of X-SDR [1], a prototype that enables the integration

[1] X-SDR is an acronym for eXtensible Suite for Dimensionality Reduction

and evaluation of any DR algorithm. X-SDR is open source and supports the evaluation of DR methods through experimentation on artificial and real world datasets. Its key features are its extensibility and user friendliness. From a user's perspective it is easy to use while from a developer's point of view, it is straightforward to extend. X-SDR supports numerous experimentation scenarios, all aligned with the evaluation methodology of applying a linear or non-linear DR method in a centralized or network environment and then assessing its quality through visualization or further experimentation with well known data mining techniques ([6]).

2 Related Work

The aim of this section is to present a brief outline of the various DR software packages. Due to space limitations we focus only on prototypes that primarily target on the evaluation and assessment of DR algorithms.

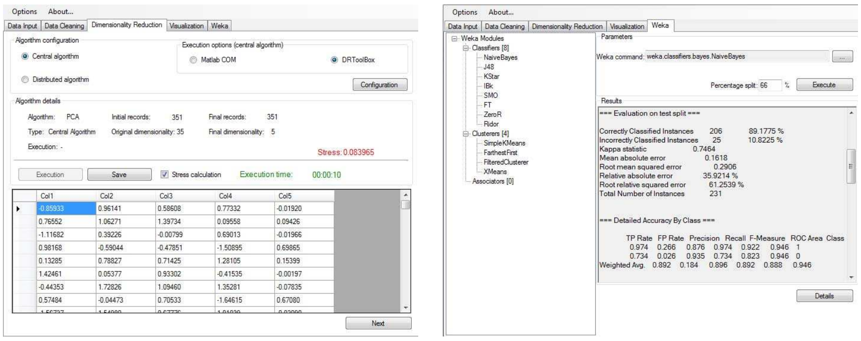
XGvis ([2],[3]) was the first attempt to offer a toolkit for experimenting with DR methods as well as visualizing their results. Unfortunately, XGvis has been confined to a number of variations of MDS ([4]). The application utilized XGobi ([9]) as a visualization frontend. XGvis evolved to GGobi ([5]), a powerful, open source suite that provides a large number of data visualization and knowledge extraction techniques specifically designed for high dimensional data. The most prominent software package in the area is the Matlab Toolbox for Dimensionality Reduction(MTDR [7]). The latter contains a vast amount of DR techniques implemented in MATLAB. Additionally, the toolbox provides implementations of methods for intrinsic dimensionality estimation, as well as functions for out-of-sample extension, prewhitening of data, and the generation of toy datasets.

Unfortunately, none of the afore described efforts assesses DR algorithms in the context of knowledge discovery. Although GGobi incorporates a large number of data mining techniques, it utilizes only a small number of DR methods. On the other hand, MTDR focuses on the algorithms rather than the evaluation of their results. Consequently, the ideal software package should combine the salient features of these two efforts and provide a user friendly frontend enabling the deployment and experimental driven assessment of any DR technique.

3 The X-SDR suite

In this section we present the extensible experimentation Suite for Dimensionality Reduction (X-SDR). We commence by outlining its software architecture followed by various implementation details. Furthermore we analyze its key aspects and highlight its novel features. Finally we provide a small experimentation scenario that highlights the merits of X-SDR.

X-SDR is structured in three layers. The first comprises the data input layer that enables interaction with various data sources ranging from simple text files to database management systems (i.e. MySQL, MS SQL Server). Its main purpose is to transform the underlying data into n-dimensional vectors. These vectors are provided as input to the DR layer which incorporates a large number of DR techniques as well as the whole MTDR suite. Additionally, it supports



Algorithm configuration

● Central algorithm

○ Distributed algorithm

Execution options (central algorithm)

○ Matlab COM ● DRToolBox

Algorithm details

Algorithm: PCA Initial records: 351 Final records: 351

Type: Central Algorithm Original dimensionality: 35 Final dimensionality: 5

Execution: -

Stress: 0.083965

☑ Stress calculation Execution time: 00:00:10

Col1	Col2	Col3	Col4	Col5
-0.85933	0.96141	0.58608	0.77332	-0.01920
0.76552	1.06271	1.39734	0.09558	0.09426
-1.11682	0.39226	-0.00799	0.69013	-0.01966
0.98168	-0.59044	-0.47851	-1.50895	0.69865
0.13285	0.78827	0.71425	1.28105	0.15399
1.42461	0.05377	0.93302	-0.41535	-0.00197
-0.44353	1.72826	1.09460	1.35281	-0.07835
0.57484	-0.04473	0.70533	-1.64615	0.67080

(a) Dimensionality reduction form

Weka command: weka.classifiers.bayes.NaiveBayes

Percentage split: 66 %

```
=== Evaluation on test split ===

Correctly Classified Instances      206          89.1775 %
Incorrectly Classified Instances     25          10.8225 %
Kappa statistic                    0.7464
Mean absolute error                  0.1618
Root mean squared error                0.2906
Relative absolute error            35.9214 %
Root relative squared error          61.2539 %
Total Number of Instances           231


=== Detailed Accuracy By Class ===

           TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
             0.974     0.266     0.876      0.974     0.922      0.946    1
             0.734     0.026     0.935      0.734     0.823      0.946    0
Weighted Avg.  0.892     0.184     0.896      0.892     0.888      0.946
```

(b) Assessment through classification

Fig. 1. Dimensionality Reduction and Data Mining forms

experimentation with distributed DR algorithms, a feature that to the best of our knowledge uniquely characterizes our framework. In order to simulate the decentralization procedure, a network profile, in the form of an adjacency matrix, should also be provided. The latter is used for the definition of a star overlay network where the central node undertakes all tasks that need to be computed centrally.

The DR layer provides two outputs. The first is a set of assessment metrics directly related to the effectiveness and efficiency of the evaluated algorithm while the second is the low dimensional embedding of the initial dataset in vector format. The assessment metrics are related to the time requirements and stress value [4] that the algorithm exhibits on a particular dataset. Moreover, for distributed algorithms, an estimation of the communication cost is also provided. The third layer aims at the assessment of the algorithms, which is accomplished either through visualization of the low dimensional dataset or further experimentation. Towards the latter we have integrated X-SDR with numerous algorithms provided by Weka [2] through proper exploitation of the OS service calls. Consequently the results of each DR algorithm are evaluated with respect to the performance of prominent clustering and classification approaches. Moreover the combination of a large number of DR methods with data mining tools promotes it as an ideal candidate for teaching as well as research activities.

The key feature of X-SDR is its extensibility which is achieved by a simple programming interface that enables any researcher to design his own algorithm and experiment through X-SDR. All required parameters are defined in XML format and incorporated as comments in the header of each MATLAB source file. Each input parameter is identified by the triplet $\{name, type, value\}$; afterwards the header of the file is parsed and the input form is dynamically created. During execution, the overarching application formulates the required calls to the MATLAB server which in turn executes the identified algorithm.

[2] Weka http://www.cs.waikato.ac.nz/ml/weka/

In order to provide a short demonstration scenario we use the 35-dimensional ionosphere dataset from the UCI repository[3]. The dataset is stored in a comma separated file and is loaded by the corresponding X-SDR module. We use the PCA [4] implementation available from MTDR and embed the dataset in a 5-dimensional space (Fig. 1(a)). The produced dataset can be visualized and even compared against the initial one. We finally evaluate the embedded dataset with the use of Naive Bayes (Fig. 1(b)) and derive a classification accuracy of 89%. The obtained value is marginally equal to the one exhibited by the same algorithm in the original space, thus concluding that PCA has successfully retained data properties while projection.

X-SDR (available through http://www.db-net.aueb.gr/panagis/X-SDR) is implemented in C♯ while all algorithms are implemented in MATLAB. The interfacing of these technologies is accomplished via the COM Automation Server. X-SDR's deployment requires MATLAB and .NET framework (version 3.5 SP1) together with Microsoft Chart Controls library.

4 Conclusion

We have presented X-SDR, an extensible experimentation suite for dimensionality reduction algorithms. X-SDR is an open source tool, integrating a large number of DR algorithms and well known knowledge discovery tool. Moreover, X-SDR enables the simulated execution of distributed DR approaches. These features promote it as an ideal candidate platform for research and teaching in academia. Future enhancements will primarily focus on incorporating distributed data mining techniques in the package thus providing an open source application for distributed knowledge discovery experimentation.

References

1. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbor meaningful? In: ICDT (1999)
2. Buja, A., Swayne., D.: Visualization methodology for multidimensional scaling. Journal of Classification 19 (2001)
3. Buja, A., Swayne, D., Littman, M., Dean, N., H.Hofmann: Xgvis: Interactive data visualization with multidimensional scaling. Technical Report (2001)
4. Carreira-Perpinan, M.A.: A review of dimension reduction techniques. Technical Report, CS-96-09, University of Shefield (1997)
5. Dianne, D.C., Swayne, D., Deborah, F.: Interactive and dynamic graphics for data analysis with r and ggobi. Journal of Computational and Graphical Statistics (2007)
6. Gabriela, H., Martin, F.: Cluster-preserving embedding of proteins. Tech. rep., Center for Discrete Mathematics and Theoretical Computer Science (1999)
7. van der Maaten, L.: An introduction to dimensionality reduction using matlab. Technical Report MICC 07-07,Maastricht University (2007)
8. Magdalinos, P., Doulkeridis, C., Vazirgiannis, M.: K-landmarks: Distributed dimensionality reduction for clustering quality maintenance. In: PKDD (2006)
9. Swayne, D., D.Cook, A.Buja: Vxgobi: Interactive dynamic data visualization in the x window system. Journal of Computational and Graphical Statistics 7 (1998)

[3] http://archive.ics.uci.edu/ml/