Modeling Stroke Diagnosis with the Use of Intelligent Techniques

No Author Given

No Institute Given

Abstract. The purpose of this work is to test the efficiency of specific intelligent classification algorithms when dealing with the domain of stroke medical diagnosis. The dataset consists of patient records of the "Acute Stroke Unit", Alexandra Hospital, Athens, Greece, describing patients suffering one of 5 different stroke types diagnosed by 127 diagnostic attributes / symptoms collected during the first hours of the emergency stroke situation as well as during the hospitalization and recovery phase of the patients. Prior to the application of the intelligent classifier the dimensionality of the dataset is further reduced using a variety of classic and state of the art dimensionality reductions techniques so as to capture the intrinsic dimensionality of the data. The results obtained indicate that the proposed methodology achieves prediction accuracy levels that are comparable to those obtained by intelligent classifiers trained on the original feature space.

1 INTRODUCTION

Stroke is a serious and rather common illness. Each year, many people all over the world suffer acute brain attacks, the vast majority of which are ischemic strokes caused by blood clots occluding brain arteries. The remainder, are hemorrhagic strokes caused by intracerebral hemorrhage and subarachnoid hemorrhage. About half of these patients are left with varying degrees of disability (those with a Rankin score of 2, 3, or 4) at 1 year after the stroke. It is this group that creates an ongoing burden to the patients themselves, to their families, and to society. In Europe stroke is one of the most important health issues as it is not only a major cause of death, but also because of the high expenses required for patient treatment. Every year in Europe about one million ischaimic strokes take place, resulting in 400,000 fatal cases [1] and the costs for each country exceed 4% of the total budget for patient treatment [2]. There is a common protocol for several hospitals of different European countries which is used since early 90s for the collection of patient stroke data from the time the patient comes to the clinic until one year after the stroke onset [3].

The importance of a timesaving and accurate method of diagnosis makes the domain a suitable candidate in applying modern approaches of intelligent computer-aided diagnosis. The handling of medical decisions concerning stroke type diagnosis by using intelligent techniques is not a new approach since a decade ago the problem had been primarily faced using inductive machine learning algorithms [4], [5], [6], [7]. The entire stroke database which we use in the present study consists of 243 diagnostic characteristics which describe 10 different types of stroke. In our experiments we used

1000 patient records of the ”Acute Stroke Unit”, Alexandra Hospital, Athens, Greece, describing patients suffering one of 5 different stroke types, diagnosed by 127 diagnostic attributes / symptoms. Reduction of the size of the database took place with the aid of medical experts. Attributes are nominal or numerical (in certain cases, part of the attributes are blank, or represent unknown or missing data, or even dont care values). The diagnoses took place in two phases, as there was a Primary Diagnosis (PD) which occurred in emergency conditions and a Final Diagnosis (FD) which came out later with the aid of laboratory examinations, within the acute stroke unit.

The aim of this study is twofold: (a) to find generalization models which can reliably differentiate among the main types of stroke, when certain diagnostic characteristics are known or can be defined, and (b) to find efficient and meaningful data representations by reducing the dimensionality of the often noisy and perhaps inaccurate initial diagnostic characteristics to reduce computational complexity and enhance the performance of the classification methods.

The paper is organized as follows: In section 2 we present the diagnostic characteristics of data collected during the first hours of the emergency stroke situation and of data related to the hospitalization and recovery phase of the patients. In section 3 we discuss the intelligent methodologies used for stroke diagnosis in this study as well as the dimensionality reduction techniques employed. In section 4 we present our experimental results and we compare them with previously derived results using similar techniques. Finally in section 5 conclusions are drawn and we explore future research potentials.

2 Stroke Registry Variables

The present study deals with the particular problem of patient classification into groups of different stroke types. Data consist of 127 decision variables which are classified into 12 diagnostic categories which can be abstracted to five (5) or even further to two (2) categories. In fact, in medical practice there exist two major categories, namely ischemic and hemorrhagic stroke. Considering their major subdivisions, 5 stroke types of interest arise. These are: *large vessel atherosclerosis, cardioembolic stroke, lacune, infarcts of unknown cause*, and *intracerebral hemorrhage*. As already mentioned above, our experimental data describe a real world problem, that is, the diagnostic characteristics of 1000 patients of the ”Acute Stroke Unit”, Alexandra Hospital, Athens, Greece. The most frequent class is Cardioembolic Stroke (33% of the total) against the other 4 classes that reach up a 15-18% each. Below we state the main characteristics encoded in the stroke database under examination.

The Primary Diagnosis (PD) dataset consists of data collected during the first hours of the emergency stroke situation, immediately after bringing the patient to the acute stroke unit. Its diagnostic characteristics are the following:

Name, Information about onset, Day of admission (day, month, year), Days of hospitalization, Indication of the place where the patient was treated, Primary Diagnosis before the first CT was obtained, Final Diagnosis on the day of discharge, Time in which the patient first presented to the hospital (emergency room), Time of admission to acute stroke unit, How did the patient come to the hospital?, Reasons of delay in considering

when the patient presented to the hospital more than 6 hours from stroke onset, Did the patient live alone?, Symptoms present on awakening?, Patient examined by another doctor at home or transferred from another hospital, Elapsed time (hrs) from stroke onset to arrival to the emergency room , Elapsed time (hrs) from arrival to the emergency room to obtaining the first CT scan, Elapsed time (hours) from stroke onset to first CT scan, Reasons for delay in obtaining the first CT scan, District areas where the patient lives, The patient's age, Sex, History of hypertension, History of diabetes, Current cigarette smoking, History of any cardiac disease, History of previous TIAs, History of elevated Hct $>$ 50%, History of elevated fibrinogen, History of migraine, Family history of coronary artery disease/stroke, History of heavy alcohol consumption, History of known lipid disorders, Oral contraceptive use, Obesity, More information about hypertension (known or unknown hypertension when the patient presented to the hospital), Duration of HTN before stroke (in years), The maximum systolic blood pressure (BP) measured in the past and reported from the patient or relatives (mmHg), Systolic Blood Pressure on admission (mmHg), Diastolic Blood Pressure on admission (mmHg), Hypertensive drugs taken by the patient the week before stroke, Duration of diabetes before stroke (in years), Current therapy of Diabetes, History of Coronary Artery Disease, Type of Coronary Artery Disease, Duration of CAD (in months), History of CABG, History of claudication, History of any valvular heart disease, History of valvular heart disease, Etiology of valvulopathy, Hospitalization for endocarditis, History of myocardiopathy, Presence of Heart Failure (HF), Duration of HF (in months), Classification of HF by NYHA, Atrial fibrillation or Sick sinus syndrome, Atrial arrhythmias on admission, Permanent vs paroxysmal arrhythmia, Known vs Unknown arrhythmia upon, Duration of arrhythmia (in months), Paced Rhythm, The patient was on coumadin or ASA in the previous week, Peripheral embolism, When did peripheral embolism occur, History of hematological disease, History of previous TIAs, Probable territory of TIAs, Elapsed time (in months) from, Duration of TIAs in minutes, Audible bruit in the neck, The main symptom or sign upon admission, Level of counsciousness upon admission, Activity at onset of stroke, Accompanying symptoms, Early course of neurological deficit, Probable location of stroke by clinical examination, Glasgow Coma Scale (3-15) (only on admission), Scandinavian Stroke Scale (0-58) (only on admission, Modified Rankin Scale (0-6) (On discharge and every follow-up), and Barthel Index (0-100).

Additional data have also been stored in the stroke database which are related to the hospitalization and recovery phase of the patients. These diagnostic characteristics are as follows:

Electrocardiogram on admission, Coronary Artery Disease by EKG, Location of MI, Hypertrophic segments seen on the EKG, Conduction abnormalities, Arrhythmias, Type of arrhythmia, Transthoracic Echocardiogram, Transthoracic Echocardiogram, Myocardial wall abnormalities, Ejection fraction (%) by TTE, Valvular heart diseases, Valvular annular calcification, Thrombus in the left ventricle, Mitral valve prolapse, Hypertrophic segments documented by TTE, Dilatated segments documented by TTE, Left atrial diameter(mm), Other findings on TTE, Transesophageal echocardiogram, Patent foramen oval, Atrial septal defect, Thrombus veget, Smoke effect, Ultrasound on the Neck, Ultrasound on the Neck, Right internal carotid, Left internal carotid, Right vertebral artery, Left vertebral artery, Cerebral angiography, Day after stroke onset

that angio was obtained, Clinical correlation, Left and right carotid distribution vessel, Left and right vertebral artery (what and where), Basilar artery, PCA1 and PCA2, Branches, CT scan or MRI that were obtained during hospitalization, LType of lesions on CT, Lacunar infarction relative to symptoms, Localization of lacunes, Localization general, Anterior vessel distribution, Posterior distribution vessel (for infarcts only), Borderzone infarcts on CT scans or MRI, Area among the major vessels of borderzone lesions, Presence in imaging studies of silent or from previous TIAs ischemic lesions, Day after stroke onset that CT2 or MRI was obtained, Hyperntense MCA sign, Early hypodensity, Presence on imaging studies of any type of Hemorragic Transformation, Edema around the lesion, Mass effect, Transtentorial herniation, Blood in ventricles, Periventricular White Matter Lesions, Volume in cm3 of lesions on images studies, Holt - 24 hour cardiac monitoring, Holter findings, Cardiac radionuclide ventriculography, Akinetic segment or aneurysm by Rad V, Ejection fraction % by Rad V, Hct (%) on admission, Platelet count on admission (X103), Urea on admission (mg/dl), Glucose on admission (mg/dl), Fibrinogen on admission (mg/dl), Prothrombin Time on admission (INR), Partial thromboplastin time on admission, Cholesterol Triglycerides, HDL, LDL in the first 2 days of hospitalization (mg/dl), Antiphospholipid and anticardiolipin antibodies, Course of disease during hospitalization, Any complication during hospitalization, What complication, The development of any cause of fever during hospitalization, Death during hospitalization or during the follow-up, Cause of death during hospitalization, Indicates recurrence, Indicates recurrence and the month after stroke onset occurred, Seizures during hospitalization or during follow-up, The most important medications administered to the patient on the first week, Medications administered to the patient at discharged, and Auxiliary therapy.

Then the database contains inclusion and exclusion criteria for thrombolysis based on specific studies. Follow-up data are also included in the stroke database. Some data were missing and some patients were lost completely on follow-up. Follow-up examination was performed within 1, 3, 6, 12, and 24 months after discharge. Some follow-up data are reported for more than 2 years, like death, cause of death, recurrence, seizures, and new heart disease. Specifically the following characteristics are included:

Follow-up examinations in the respective months, Maximum Systolic BP that was measured from another doctor, health worker or nurse, during these periods and before follow-up examination was performed.,Systolic BP that was measured, on the respective follow-up examination ,Diastolic BP that was measured e, on the respective follow-up examination ,TIA or Stroke that occurred during these particular periods and between two examinations,New Cardiac Disease that was occurred during these particular periods and between two examinations,Modified Rankin Scale that was evaluated on these follow-up examinations,Barthel Index that was evaluated on these follow-up examinations,If the patient received physiotherapy in the previous period ,Antithrombotic therapy received by thepatient particular previous period.,Antihypertensive treatment during the previous period.,Last follow-up (in case of death the last follow-up is the month of death), and Some additional information.

The Final Diagnosis (FD) dataset consists of the data collected in the last two cases above.

In order to process the two datasets (PD and FD) with the intelligent method, all features that contained unknown values were removed and the resulted datasets were further processed. Namely, all discrete (e.g. multi-class) features were decomposed into 1-to-n binary features (where n is the number of each features classes). This resulted in retaining 58 diagnostic characteristics for the final PD dataset and 25 characteristics for the final FD dataset.

3 Intelligent Methodologies for Stroke Diagnosis

3.1 Genetic Programming

One powerful search methodology of the evolutionary computation (EC) is Genetic Programming (GP) [8]. GP is widely applied in a large number of real-world problems. Extending the inherited characteristics of the EC, GP adopts a flexible variable-length solution representation and the elimination of premature convergence of the solution population. The primary GP allows for the automatic creation of expressions in mathematical, logical or algorithmic forms. As with most EC algorithms, a population of candidate solutions is commonly maintained, and successive generations are expected to enhance the solution pool (i.e. the algorithms population), enabling search into large and discontinued spaces. The unique feature of GP is a tree-like solution representation that may correspond to mathematical expressions, offering the ability to GP to perform the so-called symbolic regression. In this classification problem, the function set is comprised of logical operators, arithmetic operators and conditional IFs, adopting a GP-classification tree model [9]. As fitness measure, the total number of correct classifications was used. We have applied 10-fold cross-validation, with further use of a validation set during training, to reduce data overfitting. The GP parameters used in this application are shown in Table 1.

Table 1. GP Parameters

Parameter	Value
Population:	9000 individuals
GP implementation:	Steady-state G3P
Selection:	Tournament with elitist strategy
Tournament size:	6
Crossover Rate:	0.85
Overall Mutation Rate:	0.15
Node Mutation Rate:	0.4
Shrink Mutation Rate:	0.6
Maximum individual size:	650 nodes
Maximum generations:	100
Function set	IF greater, IF equal, IF less, And, Or, Add, Sub, Mul, PDiv
Terminal set	Number in [-1,1], Class
Overfitting measure	Validation set
Evaluation	10-fold cross-validation

3.2 Techniques for dimensionality reduction

Dimensionality reduction is the process of transforming high-dimensional data into a meaningful representation of reduced dimensionality. Dimensionality reduction is important in many domains, since it facilitates classification, visualization, and compression of high-dimensional data, by reducing the effects of the curse of dimensionality and other undesired properties of high-dimensional spaces [10].

In this study the following linear and non-linear dimensionality reduction techniques were employed: *Principal Components Analysis (PCA)* [11], *Probabilistic PCA (probPCA)* [12], *Kernel PCA* [13], *Stochastic Proximity Embedding (SPE)* [14], *Diffusion Maps (DM)* [15], [16], *Restricted Boltzmann Machines (RBM) multilayer autoencoder (AutoRBM)* [17], *Evolutionary Algorithm multilayer autoencoder (AutoEA)* [18], and *Manifold charting* [19].

Ideally, the target dimensionality should be set equal to the *intrinsic dimensionality* of the dataset which is the minimum number of parameters needed to account for the observed properties of the data [20]. In order to estimate the intrinsic dimensionality of the PD and FD datasets, the following intrinsic dimensionality estimators were employed which are based on local and global properties of the data: *Correlation dimension estimator* [21], *Nearest neighbor estimator* [21], *Maximum likelihood estimator* [22], *Eigenvalue-based estimator* [23], *Packing numbers estimator* [24], and *Geodesic Minimum Spanning Tree (GMST) estimator* [21].

4 Results and Discussion

4.1 Previously Obtained Results

Results obtained using approaches based on inductive decision trees [7], [4], [6] achieved an accuracy ranging from 83% for primary diagnosis to 86.3% for final diagnosis using 10-fold cross validation on the entire data set (850 cases were used for training and 150 cases for testing at each fold). When fuzzy modeling was attempted (time of onset, age, blood pressure, Glasgow comma scale and Scandinavian Stroke scale were among those variables modeled as fuzzy variables) a slightly higher accuracy was obtained for final diagnosis, reaching at 88.7% (a smaller subset of 38 decision variables was used for the final diagnosis experiments in this case). In both, crisp and fuzzy modeling of the problem, the highest misclassifications were generally obtained for classes 2 and 4. All the above results refer to the five-class problem. Large decision trees equivalent to more than 200 decision rules, were obtained in most experiments.

4.2 GP-results

We applied the GP algorithm to the derived PD and FD datasets and then we proceeded with the dimensionality reduction methods. All the intrinsic dimensionality estimation and dimensionality reduction task were carried out in MATLAB using the ”Matlab Toolbox for Dimensionality Reduction” [21]. For the PD dataset the dimensionality was reduced from 58 to 12 which is the intrinsic dimensionality estimate provided by the majority of all the intrinsic dimensionality estimators. For the FD diagnosis dataset

the dimensionality was reduced from 25 to 5 which is again the intrinsic dimensionality estimate provided by the majority of all the intrinsic dimensionality estimators.

Tables 2 and 3 summarize the feature dimensionality of the abovementioned data configurations. Before every GP run, we normalized the input data sets into the [-1, 1] range (using the *min-max* criterion), to facilitate the search.

Table 2. Dimensionality Reduction (PD)

PD	Features without missing values	Resulted Features	Reduced Feature Set
Binary	20	20	0
Discrete	9	35	0
Continuous	3	3	12
Total	32	58	12

Table 3. Dimensionality Reduction (FD)

FD	Features without missing values	Resulted Features	Reduced Feature Set
Binary	10	10	0
Discrete	5	15	0
Continuous	0	0	5
Total	15	25	5

Table 4 summarizes the results of the GP search, for the PD dataset. For this task, each of the data sets derived by the various dimensionality reduction methods produced lower accuracy results than the original data GP search (i.e. having available 58 features). However, it is interesting to note that this decrease in the results was relatively small for at least two cases, namely PCA (91.33%) and ProbPCA (87.86%).

Table 4. PD (Primary Diagnosis)

Feature set	10-Fold Cross Validation	Std. Dev.	Best Solution	Relative Success to Original
Original	0.5807	0.0565	0.6559	
ProbPCA	0.5102	0.0529	0.6022	87.86%
AutoRBM	0.3560	0.1156	0.5161	61.30%
AutoEA	0.3540	0.0685	0.4731	60.96%
DM	0.4631	0.0584	0.5699	79.75%
KernelPCA	0.3690	0.0791	0.5269	63.54%
SPE	0.3744	0.0590	0.5054	64.47%
PCA	0.5304	0.0463	0.6064	91.33%

Table 5. FD (Final Diagnosis)

Feature set	10-Fold Cross Validation	Std. Dev.	Best Solution	Relative Success to Original
Original	0.7701	0.0447	0.8387	
PCA	0.3851	0.0851	0.5914	50.00%
SPE	0.3743	0.0777	0.5161	48.60%
KernelPCA	0.3872	0.0877	0.5484	50.28%
DM	0.3723	0.0716	0.5054	48.34%
Manifold	0.3648	0.0733	0.4839	47.37%
AutoEA	0.3455	0.0626	0.4516	44.86%
AutoRBM	0.3669	0.0863	0.5484	47.64%
ProbPCA	0.3732	0.0726	0.4946	48.46%

Fold #7 produced the following simple and comprehensible classification tree (Figure 1), which classifies correctly 61.29% of the cases in the test set.

```
IF LAL2 = -0.10 then CL2
else (IF FEH3 > 0.06 then CL5
      else (IF COU2 > 0.06 then CL5
      else (IF AF9 < 0.80 then CL2 else CL3)))
```

Fig. 1. A derived GP classification rule-tree for the primary diagnosis

In the FD task, the reduced data sets derived by the various dimensionality reduction methods also resulted in lower success rates for the GP. This time the loss of accuracy was higher (the best reduced-data model achieved only 38.72 % accuracy in the test set whereas the original data set enabled the GP to produce a 77.01% cross-validation result in the test set). Table 5 summarizes the GP results for the final diagnosis problem.

The simple, easily interpretable, classification tree shoen in Figure 2 was derived during Fold #9, and carries 76.34% accuracy in the test set.

Overall, the GP managed to produce competitive results, deriving in some cases small and comprehensible solutions. We believe that further investigation should be performed, at least, for the PCA and ProbPCA methods in the primary diagnosis problem, since they seem promising in that task due to their ability to maintain data information in a high degree and of course due to the lower computational complexity that they induce to the GP classifier. Medical experts could apply further investigation into the resulted GP trees, in order to examine potential knowledge extraction.

```
IF LS2 < 0.57 then
   (IF LS2 < 0.58 then
     (IF MAS < 0.74 then CL2
       else (IF UN2 < 0.55 then
               (IF MAS < 0.73 then CL2
               else (IF UN2*0.57 < 0.61 then
                       (IF ARR < 0.90 then CL2 else CL3)
                     else CL3)
               else CL1)
       else CL1)
else CL5
```

Fig. 2. A derived GP classification rule-tree for the final diagnosis

Another issue that is of interest and in need of further investigation is the determinaton of the reasons why all the dimensionality reduction techniques failed to provide good results in the FD case, as compared to the PD case. A possible explanation is that usually dimensionality reduction is more meaningful when dealing with projections from very high dimensional spaces to just a few dimensions whereas the FD does not have a significantly large feature space to start with. The original feature space of only 25 attributes for the FD problem was derived, however, due to constraints imposed by the large number of missing values in the dataset. It would be therefore interesting to investigate the performance of other dimensionality reduction techniques that are able to deal with missing data such as the one described in [25].

5 Conclusion and further research

It must be obvious by now that the field of sroke medical diagnosis is very complicated. The definition of the patients' condition must be very accurate and therefore close collaboration with the expert is required. The early evaluation performed by the expert has a success rate of about 70% or even less. This fact makes it clear that a computer-based evaluation greater than 80% would be a great asset on the physician's side.

Future research of the team in this area includes the study and modeling of the error of the expert MDs between primary and final diagnosis of stroke, the effective handling of missing or "don't care" values existing among specific decision variables of the stroke database, the evaluation of the performance of more advanced classifiers and the production of useful new medical expert knowledge, which could possibly work as a guide for the efficient medical diagnosis of stroke in the future.

6 Acknowledgements

The Medical Staff of ”Alexandra” General Hospital, Athens are greatly acknowledged for their help to cope with the application domain of stroke.

References

1. Sandercock, P., Willems, H.: Medical treatment of acute ischaemic stroke. LANCET (339) (1992) 537–539
2. Carstairs, V., Gillingham, F., Mawdsley, C., Williams, A.: Resource consumption and the cost to the community. Stroke (1976) 516–528
3. Beech, R., Ratcliffe, M., Tilling, K., Wolfe, C.: Hospital services for stroke care. a european perspective, on behalf of the participants of the european study for stroke care. Stroke **27**(11) (1996) 1958–1964
4. Alexopoulos, E., Dounias, G., Vemmos, K.: Medical diagnosis of stroke using inductive machine learning. In: Proceedings of ACAI 99: Advanced Course on Artificial Intelligence, (W13) Workshop on Machine Learning in Medical Applications. (1999)
5. Nomikos, I., Dounias, G., Vemmos, K.: Comparison of alternative criteria for the evaluation of machine learning in the medical diagnosis of stroke. In: Proceedings of 3rd International Data Analysis Symposium. (1999) 63–66
6. Alexopoulos, E., Dounias, G., Vemmos, K., Nomikos, I.: Knowledge discovery & machine learning for medical diagnosis of stroke. In: 21st Annual Meeting of the Medical Decision Making Society. (1999)
7. Nomikos, I., Dounias, G., Tselentis, G., Vemmos, K.: Conventional vs. fuzzy modeling of diagnostic attributes for classifying acute stroke cases. In: ESIT-2000, European Symposium on Intelligent Techniques. (2000) 192–197
8. Koza, J.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA (1992)
9. Tsakonas, A., Dounias, G.: Hierarchical classification trees using type-constrained genetic programming. In: Proc. of 1st Intl. IEEE Symposium in Intelligent Systems. (2002)
10. Jimenez, L., Landgrebe, D.: Supervised classification in high-dimensional space: geometrical,statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man and Cybernetics **28**(1) (1997) 39–54
11. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24** (1933) 417–441
12. Tipping, M., Bishop, C.: Probabilistic prinicipal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University (1997)
13. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5) (1998) 2991319
14. Agrafiotis, D.: Stochastic proximity embedding. Journal of Computational Chemistry **24**(10) (2003) 12151221
15. Lafon, S., Lee, A.: Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9) (2006) 13931403
16. Nadler, B., Lafon, S., Coifman, R., Kevrekidis, I.: Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis: Special Issue on Diffusion Maps and Wavelets **21** (2006) 113127
17. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504507
18. Raymer, M., Punch, W., Goodman, E., Kuhn, L., Jain, A.: Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation **4** (2000) 164171
19. Brand, M.: Charting a manifold. In: Advances in Neural Information Processing Systems. Volume 15., The MIT Press (2002) 985992
20. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press Professional, Inc., San Diego,CA, USA (1990)

21. van der Maaten, L.: An introduction to dimensionality reduction using matlab. Technical Report MICC 07-07, Maastricht University, Maastricht, The Netherlands (2007)
22. Levina, E., Bickel, P.: Maximum likelihood estimation of intrinsic dimension. In: Advances in Neural Information Processing Systems. Volume 17., The MIT Press (2004)
23. Fukunaga, K., Olsen, D.: An algorithm for finding intrinsic dimensionality of data. IEEE Transactions on Computers **20** (1971) 176183
24. Kegl, B.: Intrinsic dimension estimation based on packing numbers. In: Advances in Neural Information Processing Systems. Volume 15., The MIT Press (2002) 833840
25. Kurucz, M., Benczur, A.A., Csalogany, K.: Methods for large scale svd with missing values. In: Proc. KDD-Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2007)