Research Article **Open Access**

Szymon Łukasik*, André Moitinho, Piotr A. Kowalski, António Falcão, Rita A. Ribeiro, and Piotr Kulczycki

Survey of Object-Based Data Reduction Techniques in Observational Astronomy

DOI 10.1515/phys-2016-0064
Received Sep 03, 2016; accepted Oct 26, 2016

Abstract: Dealing with astronomical observations represents one of the most challenging areas of big data analytics. Besides huge variety of data types, dynamics related to continuous data flow from multiple sources, handling enormous volumes of data is essential. This paper provides an overview of methods aimed at reducing both the number of features/attributes as well as data instances. It concentrates on data mining approaches not related to instruments and observation tools instead working on processed object-based data. The main goal of this article is to describe existing datasets on which algorithms are frequently tested, to characterize and classify available data reduction algorithms and identify promising solutions capable of addressing present and future challenges in astronomy.

Keywords: astronomy; big data; dimensionality reduction; feature extraction; data condensation

PACS: 93.85.Bc

1 Introduction

Astronomy stands on the forefront of big data analytics. In recent decades it acquired tools which have enabled unprecedented growth in generated data and consequently – information which needs to be processed. It led to the creation of two specific fields of scientific research: astrostatistics, which applies statistics to the study and analysis of astronomical data, and astroinformatics, which uses information/communications technologies to solve the big data problems faced in astronomy [58].

Since the times of individual observations with basic optical instruments astronomy transformed into a domain employing more than 1900 observatories (International Astronomical Union code list currently holds 1984 records [23]). The sizes of catalogs of astronomical objects have reached petabytes, and they may contain billions of instances described by hundreds of parameters [14]. As such, the obstacles of astronomical data analysis exemplify perfectly three main challenges of Big Data, namely volume, velocity and variety (also known as the 3Vs). Volume corresponds to both large number of instances and characteristics (features), velocity is related to dynamics of the data flow, and finally, variety stands for the broad range of data types and data sources [17].

This paper summarizes research efforts in the first of these aforementioned domains. Its goal is to present techniques aimed at alleviating problems of data dimensionality and its numerosity from a data mining perspective as well as to suggest suitable algorithms for upcoming challenges. Data is seen here as a set of astronomical objects and their properties (or their spectra). It means it is already processed from raw signals/images typically present at the instrument's level. Similarly the term "reduction" corresponds here purely to the transformation of object-based data not to the transition of raw signals/images to science ready data products. The latter can be composed of several steps and in this context data reduction could refer to several things: that raw images were processed, that photometric measurements were performed using counts stored

***Corresponding Author: Szymon Łukasik:** Faculty of Physics and Applied Computer Science, AGH University of Science and Technology; Systems Research Institute, Polish Academy of Sciences, E-mail: slukasik@agh.edu.pl
André Moitinho: CENTRA, Universidade de Lisboa, FCUL, Portugal; Email: andre@sim.ul.pt
Piotr A. Kowalski: Faculty of Physics and Applied Computer Science, AGH University of Science and Technology; Systems Research Institute, Polish Academy of Sciences; Email: pkowal@agh.edu.pl
António Falcão: Center of Technology and Systems, UNINOVA, Portugal; Email: ajf@uninova.pt
Rita A. Ribeiro: Center of Technology and Systems, UNINOVA, Portugal; Email: rar@uninova.pt
Piotr Kulczycki: Faculty of Physics and Applied Computer Science, AGH University of Science and Technology; Systems Research Institute, Polish Academy of Sciences; Email: kulpi@agh.edu.pl

Table 1: Selected sky surveys

Survey	Institution	Number of objects	Type	Time frame
Hipparcos	European Space Agency	0.12M	Optical	1989-1993
Tycho-2	European Space Agency	2.5M	Optical	1989-1993
DPOSS	Caltech	550M	Optical	1950-1990
2MASS	Univ. of Massachusetts, Caltech	300M	Near-IR	1997-2001
Gaia	European Space Agency	1000M	Optical	2013-
SDSS	Astrophysical Research Consortium	470M	Optical	2000-
LSST	LSST Corporation	4000M	Optical	2019-

in the pixels, that physical properties were extracted from spectra, etc.

In the first part of the paper, following this introduction, the scale of the data analysis problems of contemporary observational astronomy is emphasized. The second section reports on available datasets and knowledge discovery procedures. In the third section an overview of feature extraction/dimensionality reduction techniques is provided along with examples of their application for astronomical data. The fourth section is devoted to data numerosity reduction and its specific utilization for visualization of astronomical data. Both sampling and more sophisticated approaches are also addressed. Finally we suggest some existing algorithmic solutions for astronomical data reduction problems, identify future challenges in this domain, and provide concluding remarks.

2 Data Volume Problem in Observational Astronomy

The development of novel instruments used for producing astronomical data increases the data volume generated each year, at a rate which is twice that of Moore's law [46]. That is why the essence of contemporary observational astronomy could be accurately described with the metaphor of drinking water from a fire hose [49]. It reflects the fact that data processing algorithms have to deal with enormous amount of data – also on a real-time basis [58]. Consequently data reduction occurs at a low-level, at signal/image processing phase to bring down the size of transferred data. It typically involves removing noise, signatures of the atmosphere and/or instrument and other data contaminating factors. For examples of this type of reduction one could refer to [15, 16, 44, 50].

Sky surveys represent the fundamental core of astronomy. Historically, making sky observations, plotting and monitoring with the naked eye allowed significant developments to astronomical science. Today both wide-field surveys (large data sets obtained over areas of the sky that may be at least of the order of 1% of the entire Galaxy, *e.g.* see Gaia in Table 1) and deep surveys (aimed at getting important informative content from only small areas of the galaxy but with significant depth) represent keys to groundbreaking discoveries about the Universe.

Selected recent surveys frequently approached with the use of data science tools are listed in Table 1. For a more exhaustive list of astronomical surveys one can refer to [9]. It can be noticed that the number of objects listed – even for older projects – is huge. The dimensionality of the datasets depends on appropriate data preprocessing (*e.g.* frequency binning) but may reach thousands of attributes.

The extraction of knowledge from such enormous data sets is a highly complex task. Difficulties which may occur are mainly related to limits in the processing performance of computer systems – for large-sized samples – and problems exclusively connected with the analysis of multidimensional data. The latter arises mostly from a number of phenomena occurring in data sets of this type, known in literature as "the curse of multidimensionality". Above all, this includes the exponential growth in sample size, necessary to achieve appropriate effectiveness of data analysis methods with increasing dimension, as well as the vanishing difference between near and far points (norm concentration) using standard distance metrics [30].

Survey data can be explored with a variety of data science techniques. First of all, outlier detection which is aimed at identifying elements which are atypical for the whole dataset. In astronomy this technique is generally useful for discovering unusual, rare or unknown types of astronomical objects or phenomena but also for data preprocessing [59]. Another procedure is cluster analysis, which is the division of available data elements into subgroups (clusters) where the elements belonging to each cluster are similar to each other and, on the other hand,

Table 2: Selected methods of dimensionality reduction used for astronomical data

Method	Linear	Parameters	References
Principal Component Analysis	Yes	–	[27]
Kernel Principal Component Analysis	No	1	[45]
Isomap	No	1	[48]
Locally Linear Embedding	No	1	[43]
Diffusion Maps	No	2	[31]
Locality Preserving Projection	Yes	1	[20]
Laplacian Eigenmaps	No	2	[3]

there exist a significant dissimilarity between different cluster elements [33]. Identifying galaxies or groups of objects/galaxies are clustering tasks frequently performed in astronomical data analysis [13, 26]. Clustering techniques can be also used for data reduction as it will be indicated in Section 5. Both detection of outliers and clustering represent methods of unsupervised learning which are supposed to find hidden structures and relations among unlabeled data instances. Conversely, objects classification represents typical supervised learning technique. Its goal is to assign each element to one of the fixed classes, with a known set of labeled representative patterns. In astronomy it is predominantly used for identifying object types [8, 47].

Algorithms aimed at solving all the aforementioned problems are prone to negative effects from large data size, which may make their execution ineffective or even impossible. Besides applying new knowledge discovery techniques, a variety of procedures for feature extraction and data numerosity reduction can be used. They can be oriented not only towards the specific data mining task but also to data visualization which is very important for performing visual analytics on astronomical observations. These methods will be covered in more detail in the following sections.

3 Techniques of Feature Extraction

Let us assume that the object-based dataset is represented by a matrix of dimension $m \times n$:

$$X = [x_1|x_2|\ldots|x_m]^T, \quad (1)$$

with m rows representing data instances (objects) and n columns – features or attributes of all objects. The aim of reducing data dimensionality is to transform the data matrix in order to obtain its new representation with dimension $m \times N$, where N is considerably smaller than n. The reduction can be achieved either by choosing N most significant coordinates/features (*i.e.* through so called feature selection) or by means of constructing a reduced data set, based on initial features (feature extraction) [24, 57]. The latter can be treated as more general since data selection is a particularly simple case of extraction. It is important to note that any reduction procedure can be coupled with an underlying supervised learning technique – where performance of the latter is being used to evaluate the quality of the data mapping. It is common that the dimensionality of astronomical data is being reduced together with the execution of classification algorithms.

Table 2 lists the feature extraction methods commonly used for astronomical data. Besides the algorithms' names and bibliographical references, Table 2 also provides the type of mapping, *i.e.* linear/nonlinear, which states if the resulting dataset is obtained through linear transformation of the initial one. In addition, the number of required parameters – which is very important from a practical point of view – was also included. All these methods along with their applications in astronomy will be briefly presented below. Afterwards, we will also concisely present feature selection techniques.

The list of feature extraction algorithms should start with Principal Component Analysis (PCA) as it is the most commonly used dimensionality reduction method. PCA relies on orthogonal linear transformation which transforms the dataset into a new reduced feature space characterized by the greatest variance of projected data along new coordinate system axes. Practically speaking the transformation is represented by principal eigenvectors (or so called principal components) of the standardized data sample covariance matrix. PCA does not need significant computational effort and requires only one input parameter – dimensionality of reduced feature space N, which is shared by the majority of dimensionality reduction procedures. The suggested value for N, known as intrinsic dimensionality, can be estimated through the analysis of eigenvalues – it is a standard approach for establishing reduced number of features. PCA is widely used for astronomical data. As an illustration one can name the study on classi-

fication of galaxies from SDSS (Sloan Digital Sky Survey) where PCA was not only used for feature extraction but also obtaining 2D plots [36]. Besides dimensionality reduction PCA has been also used, for instance, to study the importance of features present in the Hipparcos catalog [21].

Kernel PCA constitutes an important modification of PCA by using the so called "kernel trick" [45]. Instead of principal eigenvectors of the covariance matrix Kernel PCA employs the eigenvectors of the kernel matrix. It is obtained by transforming the dataset using selected positive-semi definite kernel function K. Choice of this function can be considered as an input parameter (typically normal kernel can be used). Consequently Kernel PCA benefits from a property of constructing nonlinear mappings. It found successful applications in astronomy for supernovae photometric classification with the nearest neighbor classifier [25]. Its superiority over PCA for specific datasets was also demonstrated therein.

Isomap is a dimensionality reduction algorithm based on preserving pairwise geodesic (*i.e.* measured over the manifold) distances between data points. It estimates these distances with the shortest path between two points in the neighbourhood graph. Every data point in this graph is connected with its k neighbours, with k being an Isomap parameter. The resulting pairwise geodesic distance matrix is then transformed using classical multidimensional scaling [48]. Isomap was used for instance in classification of stellar spectral subclasses in SDSS data [5] and for discovering white dwarf + main sequence for the same survey [53]. In both cases as the classification engine Support Vector Machine method was employed, with the superiority of this solution over the one using PCA being demonstrated once more. Similar studies devoted to outlier detection have also been carried out.

Local Linear Embedding (LLE) similarly to Isomap starts with constructing a neighbourhood graph. However LLE preserves only a local geometry of the manifold surrounding each data element by representing it through a linear combination – the so-called reconstruction weights – of its k nearest neighbours (k has to be supplied as a parameter). Technically, low dimensional embedding is obtained using eigenvectors (the ones corresponding to the smallest non-zero eigenvalues) of the inner product of reconstruction weight matrix W subtracted from the identity matrix I [43]. LLE was employed for classification of objects from SDSS using their spectra in [52]. An original 1000 dimensional sample was reduced to a three dimensional subspace. As the algorithm is computationally expensive the paper also proposes a suitable data sampling scheme.

Laplacian Eigenmaps (LE) is another technique aimed at preserving local properties of the manifold. It uses additional weights corresponding to the proximity index in the set of k-nearest neighbours. It essentially means that the highest contribution to the cost function comes from the nearest neighbor. Establishing low dimensional embedding is formulated again as the eigenvalue problem through spectral graph theory [3]. Weights of the edges in the neighbourhood graph are computed using the Gaussian kernel function, therefore a supplementary parameter, *i.e.* deviation of this function σ has to be provided. Linear variant of this technique – Locality Preserving Projections (LPP) can also be named [19, 20]. While LPP has been already used with success for stellar spectral classification based on SDSS data [61] the application of Laplacian Eigenmaps for astronomical purposes was only briefly demonstrated in the paper describing new machine learning library named "megaman" [35].

Finally, Diffusion Maps (DM) rely on Markov random walk on the data represented by a graph [31]. It is based on obtaining so called diffusion distance which is related to the proximity of the data elements. The proximity is calculated during random walks performed for a limited number of time steps. The goal of dimensionality reduction is to preserve pairwise diffusion distances. The concept is derived from the field of dynamic systems. The method has been used, *e.g.* for predicting redshifts of galaxies in SDSS data by means of robust regression [40] as well as for the estimation of star formation history and supernova light curve classification [32].

It was already indicated that one alternative to data transformation is to select the most representative set of features – which is known as feature selection. It can be performed with filter methods like Relief [28] or Focus [2]. Their aim is to rank available attributes according to their informative content (or predictive power) and then select the top ones. Another approach is to use a wrapper approach. It involves iterative choice of feature subsets based on their predictive power, with forward and backward elimination being most popular procedures of this class [56]. The first starts with an empty feature set and iteratively adds useful attributes, the latter begins with the full set and in each iteration reduces it according to an optimization criterion. For more detailed description of feature selection algorithms and demonstration of their applications for astronomical data (for customized database of stars, galaxies, galactic nuclei as well as Catalina Real-Time Transient Survey and the Kepler Mission datasets) one could refer to [11, 60].

4 Methods of Instances Reduction

As previously mentioned, the data set size can be reduced to speed up data analysis calculations or make them at all feasible [7]. For astronomical datasets it is frequently used only to enable informative visualizations.

In the classical approach, data reduction is realized mostly with sampling methods [38]. Uniform sampling with or without replacement is the most widely used approach – also in astronomy. An example of its use can be found in [12]. In this study sampling was used for generating portion of data for which approximate principal components were to be obtained. Subsequent analysis concerned outlier detection for 2MASS and SDSS survey's data. In [41] automated star/galaxy clustering for digital sky data is under consideration. Randomly selected data subsets are employed for generating starting points for clustering procedures. A sample from the Digitized Palomar Sky Survey (DPOSS) is used for experimental verification. More specialized random sampling strategies related to stratified sampling – preserving the distribution of objects among classes – were also identified in the literature of the subject. For example, study [1] concerning the classification of six million unresolved photometric detections from SDSS survey obtained training data by supplementing random sample with under-represented examples. With this approach a well-known issue with random sampling, namely: poor representation of the sparsely represented examples is being alleviated.

Some specific data reduction procedures designed to be used in conjunction with individual data mining tools can be also found in the astronomical domain, as demonstrated in [54]. It uses kernel density estimation employing only a reduced, small percentage of the data sample to form probabilistic models, for instance: modeling star distribution. For that purpose the whole data set is segmented into hyper-balls with a fix radii, where each cell is associated with a kernel and a mixture weight, and subsequently the kernels are updated to fit the local distribution [54].

A variety of other methods were developed only for visualization and visual analytics. They often do not perform strict reduction, which is understood as the elimination of data elements. They simply create new data context consisting of selected data points which then may be effectively visualized. Such selection can be done manually [6], using cubes or other geometric structures [39] or based on distance from the viewpoint. More detailed review of methods dealing with large astronomical datasets only for the purpose of visualization can be found in [18].

5 Future Challenges and Suggested Algorithms

Table 1 provided a brief list of sky surveys. It included two which can be perceived as upcoming challenges: Gaia and LSST. The amount of information being generated by these project is overwhelming. LSST in one day will generate one SDSS each night for 10 years [58]. Storing data of this size and perform effective processing will not be a minor problem. It will require careful data selection and transformation aimed at enabling even simple data mining tasks.

It was already pointed out that essentially two most important features of data reduction algorithms – also in the context of forthcoming sky surveys and data generated – are required. First, scalability – that is the ability to use the same procedure even for huge datasets. It is essential to tackle datasets of ever-increasing size which we may expect in the future. The second, the low number of parameters required or their semi-automatic adjustments. Taking into account significant computational costs associated with data mining for astronomical data instances spending too much time on preliminary experiments related to data reduction should be avoided.

To reduce the number of instances we propose here to use a data condensation technique proposed by Mitra *et al.* [37]. It finds iteratively points with closest k-nearest neighbor (the distance from which is denoted by r_k) and then adds it to the reduced dataset. Simultaneously the point lying within a disc of radius $2 * r_k$ are eliminated. As the procedure requires a lot of k-NN search and range search operations using kd-trees was investigated to speed up these search operations [4]. We will demonstrate here the application of this approach for a compact version of the Hipparcos dataset with 9 features and 60876 objects. For the reduction we use only spatial coordinates of objects.

First we examined the scalability of the proposed solution. Figure 1 demonstrates that its complexity was identified to be quadratic. It means that for desktop PC used in the experiment running the algorithm for the dataset of similar structure to Hipparcos, with $m = 1000000$, $k = 5$ would take approximately 61 hours to process, which seems acceptable.

To measure the accuracy of data condensation ISE (Integrated Square Error) values were also under investigation. In general:

$$ISE(\hat{f}(x)) = \int (\hat{f}(x) - f(x))^2 dx \qquad (2)$$

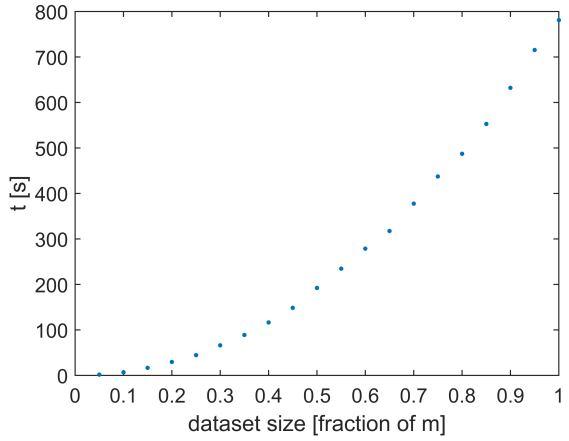

Figure 1: Scalability of Mitra *et al.* algorithm (Hipparcos dataset)

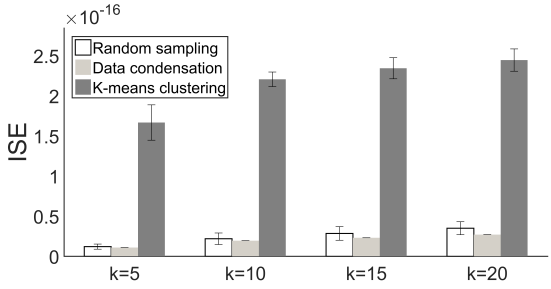

Figure 2: Integrated Square Error values obtained for probabilistic density estimates of the reduced Hipparcos data set (Hipparcos dataset)

Let us consider $f(x)$ as an original probabilistic density function. Ideally it should be of analytic form describing the whole population. Here it will be represented by an estimator obtained for the whole Hipparcos 3D sample, while $\hat{f}(x)$ will correspond to the same estimator constructed for the reduced dataset. Numerically the problem of calculating Integrated Square Error is then given by:

$$ISE(\hat{f}(x)) = \sum_{i=1}^{m} (\hat{f}(x_i) - f(x_i))^2 \tag{3}$$

with x_i being a sample element obtained from the original dataset (at the same time $m = 60876$). It basically means that we calculate an error at each sample element. We will then examine ISE in this form for three cases of data size reduction, using the same condensation intensity: random sampling (uniformly distributed), data condensation algorithm investigated here and K-means clustering (with cluster centers serving as new reduced sample elements).

Density estimates were calculated by means of a Kernel Density Estimator:

$$\hat{g}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} w_i K(\frac{x - x_i}{h}). \tag{4}$$

For approaches involving representing a group of points as one point [(2) and (3)] we use weights w_i equal to the number of points in a cluster. For the experiments a Gaussian kernel was used and smoothing parameter h was established using commonly used Silverman's "rule of thumb" [29]. As random sampling and K-means contain randomized component we used 30 replicates and report ISE mean and standard deviation. Figure 2 exhibits obtained results. It may be noticed that k-means underperforms significantly. When considering random sampling and data condensation in all cases it was the latter technique which offers better condensation quality. What is more the relative difference in ISE values of both methods grows – from 7% in case of k=5 to 26% in case of $k = 20$. For $k = 5$ results of random sampling were worse than data condensation for 22 replications of the experiment. For $k = 20$ this factor grew to 27. To conclude, the proposed approach offers reasonable time performance along with cardinality reduction which preserves important informative content of the dataset. What is more intuitive is that parameter k allows to control the intensity of reduction.

As an alternative to condensation techniques other clustering methods may be also employed (*e.g.* with elements closest to the cluster centers being preserved). The main requirements in this case are the ability to form aspherical clusters and decent computational efficiency. As an example of suitable algorithm the one demonstrated in [42] can be named.

For dimensionality reduction we are suggesting to experiment with the recent unsupervised algorithm of t-SNE. It represents an improved variant of Stochastic Neighbourhood Embedding (SNE) introduced by Hinton and Roweis [22]. In general SNE techniques start with calculating similarity matrices in both the original data space and in the low-dimensional embedding space in a way that the similarities form a probability distribution over pairs of objects [51]. The probabilities in t-SNE considered here are given by Student-t kernel computed from the input data and from the embedding. The mapping by itself is obtained by minimizing the Kullback-Leibler divergence between the two probability distributions. It was already demonstrated that t-SNE offers very-high quality mappings. For astronomical purposes the main concern could be feasibility in terms of computational time. That is why we evaluated algorithm's complexity in its Barnes-Hut variant [34] using the Hipparcos dataset. The results displayed in Figure 3 prove that it truly offers $O(m \log m)$ computational complexity as indicated in theoretical studies. It seems promising in terms of possible applications in

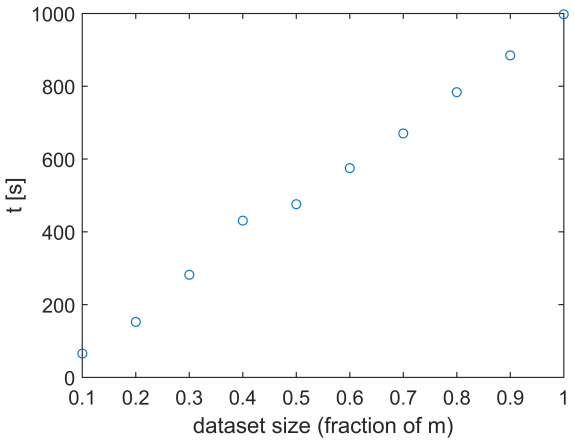

Figure 3: Scalability of Barnes-Hut t-SNE algorithm (Hipparcos dataset)

astronomy. A more exhaustive list of alternative algorithms for dimensionality reduction can be found in [10].

Finally it is worth to note that moving one step further from using specific well-performing algorithms with technical improvements (GPU and distributed computing, effective data representation etc.) is also possible. By means of alternative computing paradigm new possibilities of high-performance data mining might appear. First experiments in quantum computing for knowledge discovery prove that its a promising direction which might be used to tackle problems of future astronomical data analysis [55].

6 Conclusion

The paper studied methods of data reduction in astronomy when processed object-based data is under consideration. Besides presenting available techniques and their applications we tried to demonstrate which solutions seem more promising – also for future datasets obtained from prospective sky surveys like Gaia or LSST. The problem of discovering knowledge from astronomical datasets is not trivial – besides issues of data size difficulties related to data distribution and real-time character have to be addressed. However the benefits and the amount of useful information coming from astronomical data analysis may have a tremendous impact on space science. It can be demonstrated by the fact that the Sloan Digital Sky Survey, which has been a precursor of the field of Astroinformatics, already gave foundation to thousands of scientific publications [14]. To conclude it should be also noted that the impact of contemporary data-oriented astronomy is not limited to discovering the truth about the Universe but also about finding a way to successfully navigate through ever-present continuous streams of diverse data.

Acknowledgement: This research was supported in part by PL-Grid Infrastructure.
The contribution was co-funded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.
This work was partially funded by the Portuguese Agency "Fundação para a Ciência e a Tecnologia" (FCT) in the framework of project UID/EEA/00066/2013 and also by the European Space Agency (ESA) under contract 4000112822/14/NL/JD of project GAVIDAV.

References

[1] Abraham S. et al., A photometric catalogue of quasars and other point sources in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 2012, 419, 80-94.
[2] Almuallim H. and Dietterich T. G., Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, AAAI'91, AAAI Press, 1991, 547-552.
[3] Belkin M. and Niyogi P., Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003, 15, 1373-1396.
[4] Bentley J. L., Multidimensional binary search trees used for associative searching. *Commun. ACM*, 1975, 18(9), 509-517.
[5] Bu Y., Chen F., and Pan J., Stellar spectral subclasses classification based on Isomap and SVM. *New Astronomy*, 2014, 28, 35-43.
[6] Burgess R., Falcão A., Fernandes T., Ribeiro R. A., Gomes M., Krone-Martins A., and de Almeida A. M., Selection of large-scale 3D point cloud data using gesture recognition. In M. Luis Camarinha-Matos, A. Thais Baldissera, Giovanni Di Orio, and Francisco Marques, editors, *Technological Innovation for Cloud-Based Engineering Systems: 6th IFIP WG 5.5/SOCOLNET Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2015, Costa de Caparica, Portugal, April 13-15, 2015, Proceedings*, Springer International Publishing, 2015, 188-195.
[7] Czarnowski I. and Jedrzejowicz P., Application of agent-based simulated annealing and tabu search procedures to solving the data reduction problem. *International Journal of Applied Mathematics and Computer Science*, 2011, 21(1), 57-68.
[8] Dan G., Yan-Xia Z., and Yong-Heng Z., Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 2009, 9(2), 220.
[9] Djorgovski S. G., Mahabal A., Drake A., Graham M., and Donalek C., Sky Surveys. In T. D. Oswalt and H. E. Bond, editors, *Planets, Stars and Stellar Systems. Volume 2: Astronomical Techniques, Software and Data*, Springer, 2013, 223.
[10] Domanska D. and Łukasik S., Handling high-dimensional data in air pollution forecasting tasks. *Ecological Informatics*, 2016,

34, 70-91.

[11] Donalek C. et al., Feature selection strategies for classifying high dimensional astronomical data sets. In *Big Data, 2013 IEEE International Conference on*, 2013, 35-41.

[12] Dutta H., Giannella C., Borne K., and Kargupta H., *Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System*, SIAM, 2005, 47, 473-478.

[13] Edwards K. and Gaber M. M., *Astronomy and Big Data: A Data Clustering Approach to Identifying Uncertain Galaxy Morphology*. Springer Science & Business Media, 2014.

[14] Feigelson E. D. and Babu G. J., Big data in astronomy. *Significance*, 2012, 9, 22-25.

[15] Ferguson H. C. et al., Astronomical Data Reduction and Analysis for the Next Decade. In *astro2010: The Astronomy and Astrophysics Decadal Survey*, 2010. Position paper no 15.

[16] Freudling W. et al., Automated data reduction workflows for astronomy. The ESO Reflex environment. *Astronomy and Astrophysics*, 2013, 559, A96.

[17] Grandinetti L., Joubert G.R., and Kunze M., *Big Data and High Performance Computing*. IOS Press, 2015.

[18] Hassan A. and Fluke C. J., Scientific visualization in astronomy: Towards the petascale astronomy era. *PASA - Publications of the Astronomical Society of Australia*, 2011, 28, 150-170.

[19] He X., Cai D., Yan S., and Zhang H.J., Neighborhood preserving embedding. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, IEEE, 2005, 1208-1213.

[20] He X. and Niyogi P., Locality preserving projections. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, 2003, 153-160.

[21] Hernández-Pajares M. and Floris J., Classification of the Hipparcos input catalogue using the Kohonen network. *Monthly Notices of the Royal Astronomical Society*, 1994, 268(2), 444-450.

[22] Hinton G.E. and Roweis S.T., Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*. The MIT Press, Cambridge, 2002, 15, 833-840.

[23] IAU list of observatory codes, http://www.minorplanetcenter.net/iau/lists/ObsCodesF.html. accessed Aug 15, 2016.

[24] Inza I., Larranaga P., Etxeberria R., and Sierra B., Feature subset selection by bayesian network-based optimization. *Artificial Intelligence*, 2000, 123(1-2), 157-184.

[25] Ishida E. E. O. and de Souza R. S., Kernel PCA for Type Ia supernovae photometric classification. *Monthly Notices of the Royal Astronomical Society*, 2013, 430, 509-532.

[26] Jang W. and Hendry M., Cluster analysis of massive datasets in astronomy. *Statistics and Computing*, 2007, 17(3), 253-262.

[27] Jolliffe I.T., *Principal Component Analysis*. Springer, New York, 2002.

[28] Kira K. and Rendell L. A., The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, AAAI Press, 1992, 129-134.

[29] Kulczycki P., Kernel estimators in industrial applications. In Bhanu Prasad, editor, *Soft Computing Applications in Industry*, Springer, Berlin-Heidelberg, 2008, 69-91.

[30] Kulczycki P. and Łukasik S., An algorithm for reducing dimension and size of sample for data exploration procedures. *International Journal of Applied Mathematics and Computer Science*, 2014, 24, 133-149.

[31] Lafon S. and Lee A.B., 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*, 2006, 28(9), 1393-1403.

[32] Lee A. B. and Freeman P. E., Exploiting non-linear structure in astronomical data for improved statistical inference. In D. Eric Feigelson and Jogesh G. Babu, editors, *Statistical Challenges in Modern Astronomy V*, Springer, New York, 2012, 255-267.

[33] Łukasik S. and Kulczycki P., An algorithm for sample and data dimensionality reduction using Fast Simulated Annealing. In Jie Tang, Irwin King, Ling Chen, and Jianyong Wang, editors, *Advanced Data Mining and Applications: 7th International Conference, ADMA 2011, Beijing, China, December 17-19, 2011, Proceedings, Part I*, Springer, Berlin-Heidelberg, 2011, 152-161.

[34] Maaten van der L., Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 2014, 15, 3221-3245.

[35] McQueen J., Meila M., VanderPlas J., and Zhang Z., megaman: Manifold Learning with Millions of points. *ArXiv e-prints*, March 2016.

[36] Misra A. and Bus S. J., Artificial Neural Network Classification of Asteroids in the Sloan Digital Sky Survey. In *AAS/Division for Planetary Sciences Meeting Abstracts #40*, volume 40 of *Bulletin of the American Astronomical Society*, 2008, 508.

[37] Mitra P., Murthy C.A., and Pal S.K., Density-based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24, 734-747.

[38] Pal S. K. and Mitra P., *Pattern Recognition Algorithms for Data Mining*. CRC Press, 2004.

[39] Perkins S. et al., Scalable desktop visualisation of very large radio astronomy data cubes. *New Astronomy*, 2014, 30, 1-7.

[40] Richards J. W., Freeman P. E., Lee A. B., and Schafer C. M., Exploiting low-dimensional structure in astronomical spectra. *The Astrophysical Journal*, 2009, 691(1), 32.

[41] Rocke and Dai J., Sampling and subsampling for cluster analysis in data mining: With applications to sky survey data. *Data Mining and Knowledge Discovery*, 2003, 7(2), 215-232.

[42] Rodriguez A. and Laio A., Clustering by fast search and find of density peaks. *Science*, 2014, 344(6191), 1492-1496.

[43] Roweis S. and Saul L., Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290, 2323-2326.

[44] Schirmer M., THELI: Convenient Reduction of Optical, Near-infrared, and Mid-infrared Imaging Data. *The Astrophysical Journal Supplement Series*, 2013, 209, 21.

[45] Schölkopf B., Smola A., and Muller K.-R., Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 1998, 10, 1299-1319.

[46] Szalay A. and Gray, J., The world-wide telescope. *Science*, 2001, 293(5537), 2037-2040.

[47] Tang C.-H. et al., *Efficient Astronomical Data Classification on Large-Scale Distributed Systems*. Springer, Berlin-Heidelberg, 2010, 430-440.

[48] Tenenbaum J., de Silva V., and Langford J., A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, 290, 2319-2323.

[49] Thakar A. R., The Sloan Digital Sky Survey: Drinking from the fire hose. *Computing in Science and Engineering*, 2008, 10(1), 9-12.

[50] Valdes F. G., The Reduction of CCD Mosaic Data. In R. Gupta, H. P. Singh, and C. A. L. Bailer-Jones, editors, *Automated Data Analysis in Astronomy*, 2002, 309.

[51] van der Maaten L. and Hinton G.E., Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 2008, 9, 2579-2605.

[52] Vanderplas J. and Connolly A., Reducing the dimensionality of data: Locally Linear Embedding of Sloan Galaxy Spectra. *The Astronomical Journal*, 2009, 138(5), 1365.

[53] Wang W., Guo G., Jiang B., and Shi Y., Automatic classification for WDMS with Isomap and SVM. In *Information and Automation, 2015 IEEE International Conference on*, 2015, 1409-1413.

[54] Wang X., Tino P., Fardal M. A., Raychaudhury S., and Babul A., Fast Parzen window density estimator. In *2009 International Joint Conference on Neural Networks*, 2009, 3267-3274.

[55] Wittek P., *Quantum Machine Learning: What Quantum Computing means for Data Mining*. Academic Press, 2014.

[56] Xu L. and Zhang W.-J., Comparison of different methods for variable selection. *Analytica Chimica Acta*, 2001, 446(1-2), 475-481.

[57] Xu R. and Wunsch D.C., *Clustering*. Wiley, New Jersey, 2009.

[58] Zhang Y. and Zhao Y., Astronomy in the Big Data Era. *Data Science Journal*, 2015, 14, 1-9.

[59] Zhang Y.-X., Luo A.-L., and Zhao Y.-H., Outlier detection in astronomical data. In P. J. Quinn and A. Bridger, editors, *Optimizing Scientific Return for Astronomy through Information Technologies*, 2004, 521-529.

[60] Zheng H. and Zhang Y., Feature selection for high-dimensional data in astronomy. *Advances in Space Research*, 2008, 41(12), 1960-1964.

[61] Zhong-Bao L., Stellar spectral classification with Locality Preserving Projections and Support Vector Machine. *Journal of Astrophysics and Astronomy*, 2016, 37(2), 1-7.