Astronomical Data Analysis Software and Systems XIV **P2.1.25**
ASP Conference Series, Vol. XXX, 2005
P. L. Shopbell, M. C. Britton, and R. Ebert, eds.

Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy

Kiri L. Wagstaff

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, kiri.wagstaff@jpl.nasa.gov

Victoria G. Laidler

Computer Sciences Corporation, Space Telescope Science Institute, Baltimore, MD 21218, laidler@stsci.edu

Abstract. Modern classification and clustering techniques analyze collections of objects that are described by a set of useful features or parameters. Clustering methods group the objects in that feature space to identify distinct, well separated subsets of the data set. However, real observational data may contain missing values for some features. A "shape" feature may not be well defined for objects close to the detection limit, and objects of extreme color may be unobservable at some wavelengths.

The usual methods for handling data with missing values, such as *imputation* (estimating the missing values) or *marginalization* (deleting all objects with missing values), rely on the assumption that missing values occur by random chance. While this is a reasonable assumption in other disciplines, the fact that a value is missing in an astronomical catalog may be physically meaningful. We demonstrate a clustering analysis algorithm, KSC, that a) uses all observed values and b) does not discard the partially observed objects. KSC uses soft constraints defined by the fully observed objects to assist in the grouping of objects with missing values. We present an analysis of objects taken from the Sloan Digital Sky Survey to demonstrate how imputing the values can be misleading and why the KSC approach can produce more appropriate results.

1. Introduction

Clustering is a powerful machine learning tool that divides a set of objects into a number of distinct groups based on a problem-independent criterion, such as maximum likelihood (the EM algorithm) or minimum variance (the k-means algorithm). In astronomy, clustering has been used to analyze both images (POSS-II, Yoo et al., 1996) and spectra (IRAS, Goebel et al., 1989). Notably, the Autoclass algorithm identified a new subclass of stars based on the clustering results (Goebel et al., 1989).

However, most clustering algorithms require that all objects be fully observed: objects with missing values for one or more features cannot be clustered. This is particularly problematic when analyzing astronomical data sets, which often contain missing values due to incomplete observations or varying survey depths. Missing values are commonly handled via *imputation*, where the gap is "filled in" with an inferred value. While appropriate for some domains, this approach is not well suited to astronomical data sets, because a missing value may well be physically meaningful. For example, the Lyman break technique (Giavalisco, 2002) can identify high-redshift galaxies based on the *absence* of detectable emissions in bands corresponding to the FUV rest frame of the objects. In such cases, imputing missing values is misleading and can skew subsequent analyses of the data set.

We propose the use of a clustering approach that avoids imputation and instead fully leverages all existing observations. In this paper, we discuss our formulation of the missing data problem and expand on the KSC algorithm originally presented by Wagstaff (2004). We compare KSC analytically and empirically to other methods for dealing with missing values. As a demonstration, we analyze data from the Sloan Digital Sky Survey, which contains missing values. We find that KSC can significantly outperform data imputation methods, without producing possibly misleading "fill" values in the data.

2. Clustering Astronomical Objects with Missing Values

Missing values occur for a variety of reasons, from recording problems to instrument limitations to unfavorable observing conditions. In particular, when data is combined from multiple archives or instruments, it is virtually certain that some objects will not be present in all of the contributing sources. Little & Rubin (1987) identified three models for missing data. When values are Missing At Random (MAR, MCAR), imputation may be a reasonable approach since the values may be inferable from the observed values. The third type of missing values are Not Missing at Random (NMAR): when the value itself determines whether it is missing. This is precisely the case when objects fall below a detector's sensitivity threshold. There is no way to impute these values reliably, because they are *never* observed.

2.1. Common Methods for Dealing with Missing Values

There are three major approaches to handling missing values when clustering. The first, *marginalization*, simply removes either all features or all objects that contain missing values. The second method, *imputation*, attempts to "fill in" any missing values by inferring new values for them. The advantages and drawbacks of two marginalization and three imputation techniques are summarized in Table 1. Finally, some recent methods (Browse et al., 2003; this paper) avoid both of these approaches and instead seek to incorporate all observed values (no marginalization) without inferring the missing ones (no imputation).

2.2. Our Approach: Constrained Clustering

In our approach, we divide the data features into two categories. We use the fully observed features for clustering, and we use the partially observed features (with missing values) to generate a set of *constraints* on the clustering algorithm. We have previously shown that constraints can effectively enable clustering methods to conform to supplemental knowledge about a data set (Wagstaff et al., 2001). We use the KSC ("K-means with Soft Constraints") clustering algorithm, proposed by Wagstaff (2004), to incorporate information from the partially observed

Table 1. Comparison of marginalization and imputation methods.

Technique	**Advantages**	**Drawbacks**
Feature Marginalization: Omit features with missing values	Simple	Lose information about all objects
Object Marginalization: Omit objects with missing values	Simple	Lose objects
Mean Imputation: Replace each missing value with data set mean	Simple	Likely to be inaccurate; mean value may never truly occur
Probabilistic Imputation: Replace with random value according to data set distribution of values	Inferred values are "real" (actual observations)	Inferred values may have no connection to the objects
Nearest Neighbor Imputation: Replace with value(s) from the nearest neighbor	Inferred values are "best possible guess"	Inferred values may still be inappropriate (unobservable)

features as a source of information that supplements the fully observed features. Before discussing our experimental results, we will describe what we mean by constraints and briefly outline the KSC algorithm.

A *soft constraint* between two objects indicates a preference for, or against, their assignment to the same cluster. We represent a soft constraint between objects o_i and o_j as a triple: $\langle o_i, o_j, s\rangle$. The strength, s, defines the nature and confidence of the constraint. A positive value for s indicates a preference towards clustering o_i and o_j together; a negative value suggests that they should be assigned to different clusters.

The KSC algorithm is based on the basic k-means algorithm first proposed by MacQueen (1967). While the k-means algorithm seeks to minimize the total variance, V, of a proposed partition P, KSC minimizes the combination of the variance and a penalty, CV, for constraint violations:

$$f(P) = (1-w)\frac{V(P)}{V_{max}} + w\frac{CV(P)}{CV_{max}} \tag{1}$$

The CV penalty is calculated as the sum of the squared strengths, s, of all constraints violated by the partition P. The quantities V and CV are normalized by their maximum possible values. The user-specified weight parameter w indicates the relative importance of variance versus constraint violations; a good value can be chosen based on performance on a small labeled subset of the data.

3. Experimental Results

The Sloan Digital Sky Survey (SDSS) contains observations of 141 million galaxies, quasars, and stars (as of data release 3), many with missing values. We selected a small subset of this data for our experiments. The features we used were brightness (psfCounts), texture, size (petroRad), and shape (M_e1 and M_e2). 42 of the 1507 objects we analyzed have missing shape features. We clustered this data set based on the three fully observed features and generated constraints based on all of the features. That is, for each pair of ob-

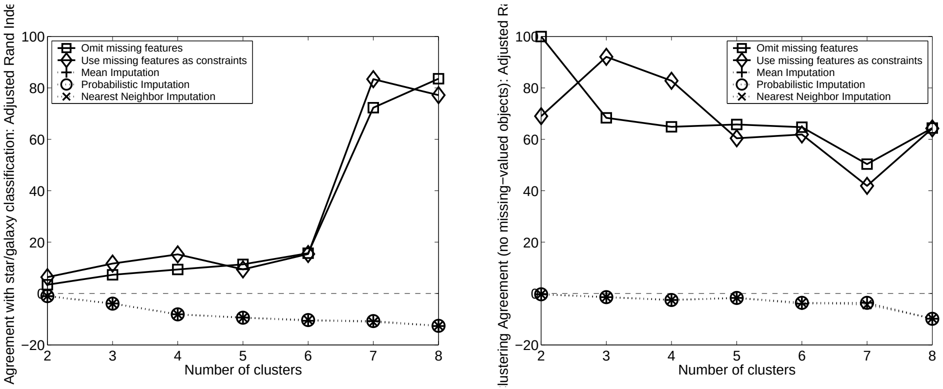

Figure 1. Empirical comparison of imputation, marginalization, and KSC methods. Figure (a) shows agreement with true star/galaxy separation; figure (b) shows the impact (changes in cluster assignment) for the fully observed objects. The dashed line shows the agreement expected by random chance.

jects o_1, o_2 that had observed shape features, we generated a constraint with strength $s = \sqrt{\sum_i (f_i(o_1) - f_i(o_2))^2}$. The result was 1,072,380 constraints that were "mined" from the data set.

3.1. Separating Stars and Galaxies

In our experiments, we calculated the agreement between the partition created by each method and the "true" star/galaxy classification (based on SDSS labels), using the Adjusted Rand Index (ARI), which was proposed by Hubert and Arabie (1985). An ARI of 0 indicates the amount of agreement expected by randomly assigning the same number of items to the specified number of clusters, with the same number of items per cluster.

Figure 1(a) shows performance results for partitions that contain two to eight clusters. All three imputation methods negatively impact performance, producing results that are actually worse than that expected by random chance. We observe that when the shape features are completely ignored (marginalization), agreement steadily increases as the number of clusters goes up; the clustering method is able to assign unusual objects to their own clusters and more accurately separate stars and galaxies in the rest. Including the observed shape information as constraints is competitive with, and sometimes superior to, marginalization. We expect that if shape information were more relevant for separating stars from galaxies, then higher agreement would be observed.

3.2. Impact on Fully Observed Objects

In Figure 1(b), we assess the clustering impact on the fully observed objects in the data set. Since only 42 of the 1507 objects have missing values, their inclusion should result in little change in the cluster assignments for the fully observed objects. We find that this is true for marginalization and for KSC (agreement for the fully observed objects stays high), but again the imputation methods do not perform well. Imputing the missing values seems to significantly impact the placement of other objects in the data set.

4. Conclusions

In this paper, we have demonstrated that the imputation methods commonly used to cluster objects when some feature values are missing can be particularly misleading when applied to astronomical data. Our empirical results with SDSS data show that imputation methods prevent the correct separation of stars and galaxies, while KSC with constraints generated from the fully observed objects performs much better (up to 90% improvement). In addition, KSC minimizes the impact on cluster assignments for the fully observed objects.

When there are a only a few observed values for a given feature, imputation methods are even less reliable because they have too little information from which to infer the missing values. This is exactly the case where KSC is most efficient, and effective, since the runtime require to generate and enforce all of the constraints is proportional to the number of fully observed objects. We plan to explore this further in future experiments.

Acknowledgments. We would like to thank Marie desJardins for her suggestions. This work was supported in part by NSF grant IIS-0325329. Funding for the Sloan Digital Sky Survey (SDSS) has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/.

References

Browse, R. A., Skillicorn, D. B., & McConnell, S. M. (2003). Using competitive learning to handle missing values in astrophysical datasets. *Workshop on Mining Scientific and Engineering Datasets, SIAM Data Mining Conference.*

Giavalisco, M. (2002). Lyman-break galaxies. *Annual Reviews of Astronomy & Astrophysics*, *40*, 579–641.

Goebel, J., Volk, K., Walker, H., Gerbault, F., Cheeseman, P., Self, M., Stutz, J., & Taylor, W. (1989). A Bayesian classification of the IRAS LRS Atlas. *Astronomy and Astrophysics*, *222*, L5–L8.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Little, R. J. A., & Rubin, D. A. (1987). *Statistical analysis with missing data*. John Wiley and Sons.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Symposium on Math, Statistics, and Probability* (pp. 281–297). Berkeley, CA: University of California Press.

Wagstaff, K. (2004). Clustering with missing values: No imputation required. *Classification, Clustering, and Data Mining Applications (Proceedings of the Meeting of the International Federation of Classification Societies)* (pp. 649–658). Springer.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained k-means clustering with background knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 577–584). Morgan Kaufmann.

Yoo, J., Gray, A., Roden, J., Fayyad, U. M., de Carvalho, R. R., & Djorgovski, S. G. (1996). Analysis of digital POSS-II catalogs using hierarchical unsupervised learning algorithms. *Astronomical Data Analysis Software and Systems V* (pp. 41–44).