# Feature Selection in Scientific Applications

Erick Cantú-Paz
cantupaz@llnl.gov

Shawn Newsam
newsam1@llnl.gov

Chandrika Kamath
kamath2@llnl.gov

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
7000 East Avenue, L-561
Livermore, CA 94550

## ABSTRACT

Numerous applications of data mining to scientific data involve the induction of a classification model. In many cases, the collection of data is not performed with this task in mind, and therefore, the data might contain irrelevant or redundant features that affect negatively the accuracy of the induction algorithms. The size and dimensionality of typical scientific data make it difficult to use any available domain information to identify features that discriminate between the classes of interest. Similarly, exploratory data analysis techniques have limitations on the amount and dimensionality of the data they can process effectively. In this paper, we describe applications of efficient feature selection methods to data sets from astronomy, plasma physics, and remote sensing. We use variations of recently proposed filter methods as well as traditional wrapper approaches, where practical. We discuss the general challenges of feature selection in scientific datasets, the strategies for success that were common among our diverse applications, and the lessons learned in solving these problems.

## Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—feature evaluation and selection.

## General Terms

Algorithms, Experimentation, Measurement

## Keywords

Feature selection, data mining, astronomical survey, plasma physics, remote sensing.

## 1. INTRODUCTION

Scientific data sets generated by computer simulations, observations, or experiments present challenges. For instance, many scientific applications require the extraction of features from low-level data, such as images or mesh data from simulations. The data can be noisy, especially if coming from experiments or sensors, and removing the noise without affecting the signal is difficult. In contrast with commercial data, assigning labels to scientific data usually requires a domain scientist to identify the objects of interest. Besides being tedious, this subjective process is prone to errors and experts often disagree on the labeling. Another difficulty is that scientific data is sometimes obtained from different sources and is captured at different resolutions or with different instruments, so data fusion becomes necessary to incorporate all the data into the analyses.

In this paper we are concerned with the problem of feature selection. This problem has its origins in some of the difficulties mentioned above. In particular, there are many methods to extract features from low-level image or simulation data that range in sophistication from simple statistics of the variables of interest to shape or texture descriptors. It is likely that the collection of data was not performed with a particular analysis in mind. Therefore, the data may contain irrelevant or redundant features that affect the analysis negatively. Domain information is very helpful in pruning the data and in identifying candidate variables, but in many cases the size and dimensionality of scientific data make it difficult to use available domain information to identify features that discriminate between the classes of interest.

Regardless of these difficulties, data mining is gaining acceptance in many scientific fields. This paper describes applications of feature selection to three very different scientific data sets. The number of features in these applications vary from a few tens to a few hundreds. We describe the challenges common to these applications, the strategies we followed to face these challenges, and the general lessons we learned in solving these problems.

The next section describes the feature selection algorithms that we use. Section 3 describes a classical classification problem with the goal of building a predictor to identify galaxies of a particular type. We describe how the data was generated with input from astronomers, why we assumed that feature selection was necessary, and the results using different automatic methods as well as further manual reductions. The second application is described in Section 4. The goal is to identify which variables might explain the presence of a harmonic oscillation on the edge of plasma in fusion experiments. Section 5 discusses the problem of detecting human settlements in satellite imagery. These data contain the highest number of features of the problems we considered, and effective feature selection could save con-


*KDD'04,* August 22–25, 2004, Seattle, Washington, USA.

siderable computing resources used in creating and storing these features. Finally, Section 6 summarizes the common approaches, the lessons learned, and the conclusions.

## 2. FEATURE SELECTION ALGORITHMS

The feature selection problem has received considerable attention and numerous feature selection algorithms have been proposed. In this paper, we use four filters, two classical wrappers, and one hybrid filter-wrapper method. We use the filters and the filter-wrapper hybrid to rank the features. We evaluate the rankings using a naive Bayes classifier on increasingly larger subsets of the ranked features and report the 10-fold crossvalidation estimate of the prediction error. We also report results of a decision tree, but only when they are better than the naive Bayes.

We introduce into each dataset a "sentinel" random feature that is uniformly distributed in the interval [0,1]. Features ranked lower than the sentinel should be discarded.

### 2.1 Filters

The first filter that we use estimates how well a feature separates the data into different classes using the Kullback-Leibler (KL) distance between histograms of feature values. For each feature, there is one histogram for each class. Numeric features are discretized using $b = \sqrt{|D|}/2$ equally-spaced bins, where $|D|$ is the size of the training data. The histograms are normalized by dividing each bin count by the total number of elements to estimate the probability that the $j$-th feature takes a value in the $i$-th bin of the histogram, given a class $n$: $p_j(d = i|n)$. For each feature $j$, we calculate the class separability as

$$\Delta_j = \sum_{m=1}^{c} \sum_{n=1}^{c} \delta_j(m, n), \tag{1}$$

$c$ is the number of classes and $\delta_j(m, n)$ is the KL distance between histograms corresponding to classes $m$ and $n$. The features are ranked by sorting them in descending order of the distances $\Delta_j$.

The second filter ranks features by sorting them in descending order of Chi-square statistics computed from their contingency tables. The contingency tables have one row for every class and the columns correspond to possible values of the feature [7]. Numeric features are represented by histograms, so the columns of the contingency table are the histogram bins. The Chi-square statistic for feature $j$ is

$$\chi_j^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

where the sum is over all the cells in the contingency table, $o_i$ stands for the observed value, and $e_i$ is the expected frequency of items.

We can also rank features based on criteria used by decision trees in deciding which variables to use in splitting the data. Decision trees examine one feature at a time and find the split that optimizes an impurity measure. In this paper, we measure the impurity with the Gini index [2] and rank the features according to the impurity of their optimal split. This filter provides information on the optimal thresholds on the values of the features, which may be of interest to the scientists. We refer to this method as a stump filter.

For the last filter, we adopted a method suggested by Mardia et al. [10] to use PCA to eliminate unimportant variables. Starting with the eigenvector that corresponds to the smallest eigenvalue of the covariance matrix, we discarded the variable with the largest coefficient (in absolute value) in that vector. We then proceed to the next eigenvector and discarded the variable with the largest coefficient, among the variables not discarded earlier. We continued with this process until all variables were ranked.

### 2.2 Wrappers

Sequential forward selection (SFS) and sequential backward elimination (SBE) are two classic greedy wrappers. Forward selection starts with an empty set of features. In the first iteration, the algorithm considers all feature subsets with only one feature. The feature subset with the highest accuracy is used as the basis for the next iteration. In each iteration, the algorithm tentatively adds to the basis each feature not previously selected and retains the feature subset that results in the highest estimated performance. The search terminates after the accuracy of the current subset cannot be improved by adding any other feature. Backward elimination works in an analogous way, starting from the full set of features and tentatively deleting features.

In these algorithms, each feature subset is evaluated by estimating the accuracy of a classification algorithm with the candidate subset of features. In this paper, we use 10-fold crossvalidation to estimate the accuracy.

### 2.3 Boosting Filter-Wrapper Hybrid

This algorithm is a generalization of Das' filter-wrapper hybrid algorithm [3]. In each iteration, with a user-supplied filter, the algorithm ranks the features that have not been selected so far and adds the highest-ranking feature to the feature subset. For this ranking we use the KL class separability filter described above. Then, a "re-weighting" classifier is trained using the current feature subset and classifies the instances in the training set. The weights of the instances are updated using the standard AdaBoost procedure [5] (giving more weight to instances misclassified), and the algorithm iterates.

We follow Das and train the re-weighting classifier ignoring the weights of the instances. The weights are used only by the filter to rank the unselected features in each iteration. The algorithm stops when the accuracy of a "stopping" classifier trained with all the selected features does not improve from the previous iteration. Das argued (and we confirmed) that using the accuracy on the training set was adequate for stopping the algorithm. In our experiments we use naive Bayes as the "re-weighting" and "stopping" classifiers. Preliminary tests did not show large differences in the error rates of the final classifiers when stumps were used to re-weight instances and trees were used to stop the algorithm.

## 3. FIRST ASTRONOMICAL SURVEY

The first data set that we examine comes from the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey [1]. Our goal is to identify galaxies with a bent-double morphology as they indicate the presence of clusters of galaxies, a key project within the FIRST survey. Scientists currently identify the bent-double galaxies visually, which—besides being subjective, prone to error and tedious—is becoming increasingly infeasible as the survey grows.

Data from FIRST are available at sundog.stsci.edu as image maps and a catalog. The images in figure 1 are close-

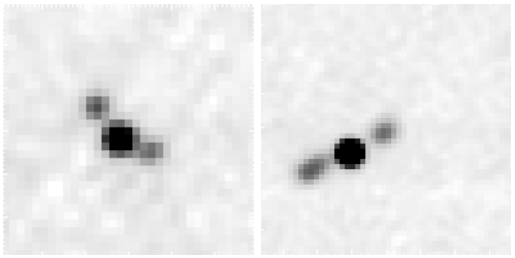

**Figure 1: Examples of bent-double (left) and non-bent double (right) radio galaxies.**

ups of bent-double and non-bent double galaxies. The astronomers obtained the catalog by fitting two-dimensional Gaussians to each radio source. Each entry in the catalog corresponds to a Gaussian and includes information such as the location and size of the Gaussian and the peak flux.

We identified candidate features for this problem through extensive conversations with FIRST astronomers, who placed great importance on spatial features such as distances and angles. In total, we extracted 99 non-housekeeping features, which are described in detail elsewhere [4].

Our training set is relatively small, containing 195 examples. Since the bent- and non-bent-doubles must be manually labeled by FIRST scientists, putting together an adequate training set is non-trivial. Moreover, the labeling of galaxies is subjective and the astronomers often disagree in the hard-to-classify cases. There is also no ground truth we can use to verify our results. These issues imply that the training set itself is not very accurate, and there is a limit to the accuracy we can obtain. Even with extensive domain knowledge, reducing the number of features in this problem was problematic. For example, in consultation with the astronomers, we generated three different measures of symmetry and of "bentness." These measures are clearly correlated, but it is not obvious which one(s) should be preferred.

Figure 2 presents 10-fold crossvalidation estimates of the error rates of a naive Bayes classifier using increasingly large feature subsets. The PCA and the stump filters find small feature subsets that result in the lowest accuracy (both reach 12.1%). Interestingly, as we test larger feature subsets identified by the PCA ranking, the error of the classifier becomes the worst. The error using all the features is 17.3%, so the observed improvements are significant (according to a two-sided $t$-test at 0.05 level of significance).

This data is sufficiently small for SFS and SBE wrappers to be practical. SFS found a subset with four features that resulted in an error rate of 12.63%, not significantly different from the best result of the ranking methods. On the other hand, SBE did not eliminate any variables from this data. Of the four features that SFS found, one is the total area of the three Gaussians that represent the galaxy. This feature is irrelevant to the identification of bent-double galaxies (because the area depends on the proximity of the galaxy to the Earth, not of any intrinsic property of the galaxy) and can be eliminated.

Our previous experience with this data suggested that the best accuracies are usually achieved using features extracted considering triplets of catalog entries (as opposed to pairs or single entries) [4]. There are only 20 of these features, and the results are presented in the right panel of figure 2. In this case, the Chi-square and the KL distance filters found subsets of 10 and 11 features that resulted in the lowest errors (8.9 and 10.5%, respectively). These errors are significantly different from the best results obtained using all the features. SFS found a subset of three features (again including the total area), that resulted in an error of 10%. SBE failed again to eliminate any features.

Although the goal of this project was to produce a predictor to classify galaxies as accurately as possible, it is important to examine the composition of the feature subsets selected. As we have noted above, SBE produces impressively small feature subsets that always include one obviously irrelevant feature. This is possible since the naive Bayes is insensitive to truly irrelevant features, but stochastic errors of the crossvalidation estimates may make an irrelevant feature appear as giving a small advantage.

Except for PCA, the filters rank highly features related to symmetries and angles, which are features the astronomers and we expected. PCA selects features that, although unexpected, appear to have good discriminatory power, which we confirmed by a simple exploratory data analysis observing box-plots and histograms.

## 4. FUSION DATA

Sometimes the goal in a scientific application is not to build a predictor, but to discover a set of features that may provide leads into the problem of interest to the scientist. We present an application on fusion physics data where our goal is to identify which candidate variables are related with an interesting state of the plasma.

Fusion is a nuclear reaction where lighter elements combine to form a heavier element. This reaction releases large amounts of energy that, if harnessed and controlled, represent a sustainable and environmentally sound energy source. The most successful and promising fusion confinement device is known as a tokamak. A high-confinement mode (H-mode) of operation is the choice for next generation tokamaks, but it comes at a significant cost due to effects of edge localized modes (ELMs). ELMs cause rapid erosion of some components in tokamaks and giant ELMs can destroy other critical components. Recently, a "quiescent H-mode" of operation has been observed in the DIII-D tokamak operated by General Atomics. Quiescent operation is important because there are no ELMs. Scientists have detected that the presence of an edge harmonic oscillation (EHO) is associated with the QH-mode.

EHOs are identified mostly by visual means using the Fourier spectrum of data from a magnetic probe. Data from other sensors are consulted to verify the existence of the EHO. A program developed at General Atomics implements the rules scientists have derived to identify EHOs. Although the program *identifies* EHOs satisfactorily, it does not *explain* their presence.

Scientists are interested in knowing which variables are related to the appearance of EHOs. Our approach to this problem is to identify which of the candidate variables are relevant to create models that predict the EHOness of the experimental data.

Each experiment in DIII-D lasts for approximately six seconds and data from numerous sensors are recorded and stored in a database. With input from a domain scientist, we

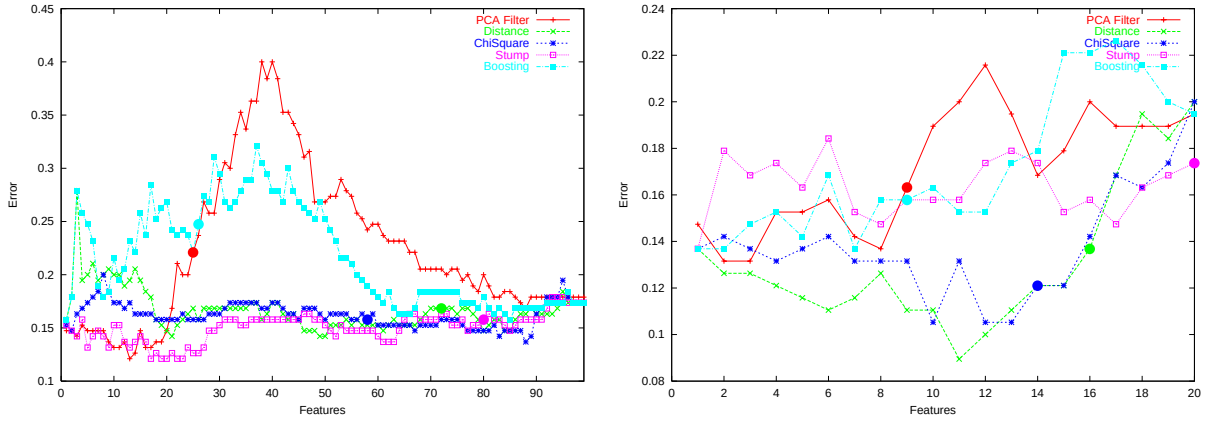


**Figure 2: Error rates varying the number of features using FIRST data. The left graph shows results considering all 99 features and the right graph shows results considering only the 20 triplet features. The large dots represent the rank of the random noise feature.**

extracted the values of 37 candidate variables that describe approximately 700 experiments. Each 50 ms time window of each experiment receives a binary label (high/low EHOness) from the program that detects EHOs.

The data needs preprocessing before being input to the feature selection algorithms. One of the major difficulties is that the data from multiple sensors is sampled at different rates or may start or end at different points in time. This is a typical problem with scientific data and requires that the data be registered. For a variety of reasons, some sensors may not have been activated for an experiment, and in consultation with our collaborator, we decided to discard the time windows that contain at least one missing value.

The size and dimensionality of the data still allows for a meaningful exploratory data analysis. Visual examination of box-plots and histograms revealed that the data contained many outliers. Using the median value of each variable in each time window eliminated some outliers, but since many still remained, we decided to eliminate the time windows that contained at least one variable in the top or bottom percentile of its range.

As we saw in the previous example with FIRST data, labeling the data manually usually means that the training sets are small. However, the fusion data does not suffer from the typical lack of labeled data, because the labeling is performed by a program. After the preprocessing, our training set consists of 41818 instances, and it is easy to expand it to include additional experiments.

Figure 3 presents the error rates of a naive Bayes classifier trained on increasingly large feature subsets. As with the FIRST data, the PCA filter produced a compact feature subset that results in the lowest classification error of 17.3%. Although this error is not notably smaller than the error obtained with all the features (20.9%), the fact that very few features are necessary to explain the presence of EHOs is interesting.

There is significant overlap between the top ten features ranked by the different methods, except for the PCA filter that selects features that the other methods rank lower. Six features were ranked in the top ten by four filters, and an additional two were ranked in the top ten by three filters.

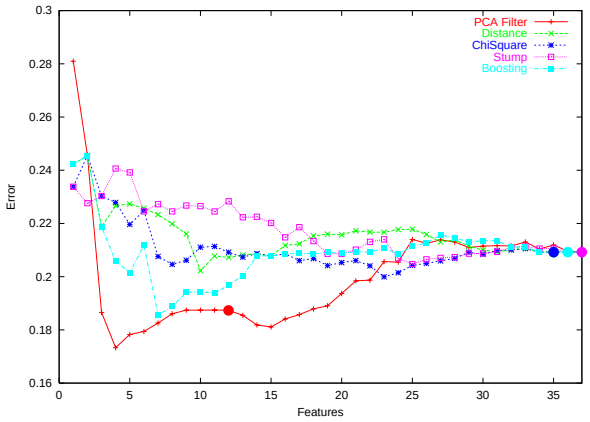


**Figure 3: Error rates varying the number of features using the fusion data. The large dots represent the rankings of the random noise feature.**

SFS and SBE found feature subsets with three and four features, respectively, and both subsets resulted in accuracies of 16.2%. There was only one common feature in these subsets, and it was a feature that appeared consistently in the top ten rankings with the filters (except in the PCA).

Interpretation of the physical significance of these results is beyond our abilities. The goal of this project was to identify variables that the scientists can use as leads to explain the presence of EHOs. While these results might provide useful leads, this is an ongoing collaboration with physicists where we are exploring other feature selection methods as well as summarization of the results in novel ways.

## 5. REMOTE SENSING DATA

In this application, we consider the use of data mining techniques to automate the identification of human settlements in satellite imagery. This is an essential step in the production of maps of human settlements, which are used in studies of urbanization, population movement, etc. We

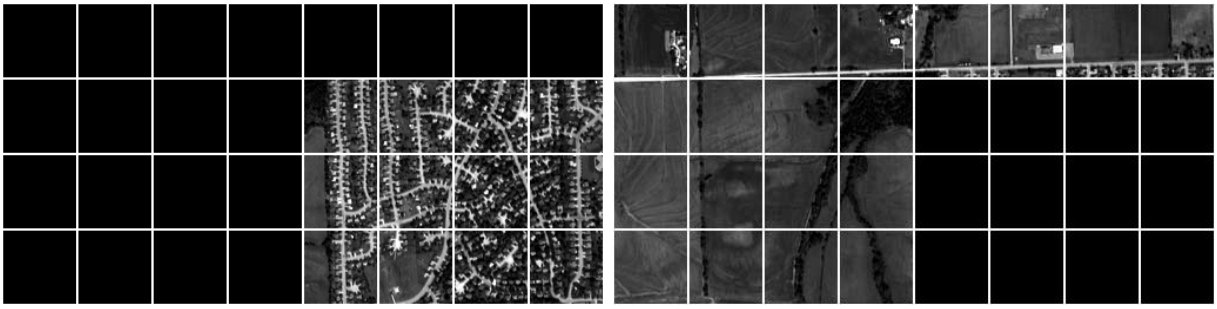

**Figure 4: Example of a region in satellite imagery illustrating the ground truth with inhabited tiles (left) and uninhabited tiles (right). Original satellite image by Space Imaging.**

used 4-band multi-spectral (near-infrared, red, green, and blue) images at 4m ground sample distance from IKONOS that are available at www.spaceimaging.com.

We considered two satellite images—one from a region in Nebraska and the other from northern Mexico. We first divided each image into non-overlapping tiles, each of size $64 \times 64$ pixels. Next, we extracted image texture features from each of the four spectral bands in the tiles. We manually labeled each tile as either being inhabited—that is, being predominantly composed of man-made structures—or as being uninhabited. Figure 4 shows examples of both types of tiles. We removed tiles that were of uniform intensity. For our images this resulted in 7419 instances.

A number of texture features have been proposed over the last several decades, but none has proven superior for all applications, so deciding on the appropriate features remains a challenge. This challenge is compounded in the analysis of multi-band imagery, since it is also not clear which band(s) should be used to compute the texture features.

**Table 1: Minimum error rates using the naive Bayes for different feature selection methods with features considered independently and in combination.**

| Method | Pow. Sp | GLCM | Wavelet | Gabor | All |
|---|---|---|---|---|---|
| No filter | 29.0 | 27.5 | 28.8 | 33.2 | 41.8 |
| PCA | 28.0 | 27.1 | 28.2 | 27.8 | 28.8 |
| KL | 27.1 | 26.0 | 26.7 | 26.1 | 26.0 |
| $\chi^2$ | 27.0 | 26.0 | 26.5 | 26.1 | 26.0 |
| Stump | 28.2 | 26.6 | 27.8 | 26.6 | 26.9 |

**Table 2: Minimum error rates using the decision tree for different feature selection methods with features considered independently and in combination.**

| Method | Pow. Sp | GLCM | Wavelet | Gabor | All |
|---|---|---|---|---|---|
| No filter | 26.5 | 25.1 | 24.6 | 25.8 | 25.6 |
| PCA | 25.9 | 24.3 | 24.3 | 25.4 | 24.8 |
| KL | 25.9 | 24.5 | 24.6 | 25.5 | 25.1 |
| $\chi^2$ | 25.7 | 24.7 | 24.6 | 25.6 | 24.9 |
| Stump | 25.4 | 24.3 | 24.4 | 25.4 | 24.8 |

We extracted texture features based on (1) gray level co-occurrence matrices (GLCMs) [6], (2) the Fourier power spectrum [11], (3) wavelets [8], and (4) Gabor filters [9]. Each of the four sets of texture features was extracted from each of the four spectral bands resulting in a total of 496 texture feature components. More details are available elsewhere [11]. We performed two sets of experiments. First, we considered the four sets of texture features independently, and then we combined all the features. Our goal was to understand the performance of the feature selection algorithms, identify if any of the features performed better than the rest, and see if combinations of features worked better than each set considered independently.

Tables 1 and 2 summarize the minimum error rates for each feature selection method using each of the four sets of texture features in isolation and in combination, with the naive Bayes and decision tree classifiers, respectively. For this problem, we found that decision trees gave better results than the naive Bayes classifier, often by more than 1% error rate. When the combination of the four sets of texture features was used without any feature selection, the naive Bayes classifier had an error rate of 41.8% compared to the decision tree error rate of 25.6%. This is to be expected as the naive Bayes classifier is known not to perform well in the presence of many features and the decision tree can be considered to have in-built feature selection. Thus, the explicit use of feature selection benefits the naive Bayes classifier more than the decision tree classifier. Similarly, we observe that when we consider only the Gabor features, which are more numerous than the other features, the error rate is higher for the naive Bayes classifier (33.2%) than the decision trees (25.8%).

The PCA filter selected texture features and spectral bands that were rarely selected by other methods. Further, it ranked the noise feature quite highly (often within the top 10%), even though this feature was irrelevant. We believe this poor performance is the result of the PCA filter ignoring the class of each instance.

We found that there was not much difference in performance among the remainder of the feature selection techniques, though with the naive Bayes classifier, the PCA filter and the stump filter selected features that performed slightly worse than other techniques.

By examining the top ten selected features, we made several interesting observations that were consistent across the two sets of experiments. We observed that features of the green and near-infrared bands were selected more often than features of the blue and red bands. This suggests that we

could reduce the computation time and storage by focusing on only two bands. Also, a majority of the top ten features are from the GLCM category, while the wavelet and Gabor features are selected less frequently. Power spectrum features are rarely selected, agreeing with our prior experience with these data [11]. The GLCM features selected most often in the top ten are entropy and inverse difference moment. Other GLCM features were selected rarely or occasionally, suggesting that we may be able to reduce the number of GLCM features to two or three.

The wavelet and Gabor features selected in the top ten correspond to the higher frequencies. While we considered three levels in the wavelet decomposition, only the first two were ever selected. Similarly, for the Gabor features, only the two highest of five scales were selected. Further, for the Gabor and wavelet features, it was mainly the energy feature that was selected in the top ten; the standard deviation was rarely selected. These observations can again be used to reduce substantially both the computation time and the storage requirements.

## 6. DISCUSSION

The three diverse examples of scientific applications that we presented in this paper illustrate the difficulties of performing feature selection in scientific data.

The labeling of examples is a common source of problems. In FIRST and the human settlements problems, the labeling was performed manually. This limits the size of the training set, and together with the error-prone and subjective nature of the labeling, restricts the accuracy we can expect from classification methods. In the fusion data, the labeling is automated, but it may also contain mistakes, as the labeling program implements heuristics that may not be always valid.

The three applications demonstrated that preprocessing is crucial for the success of these projects. In scientific data, we frequently generate features from low-level data that may be noisy. In the FIRST data, the noisy images were processed into a catalog (by the astronomers) creating fairly noiseless processed data that we used to create high-level features; in the fusion data we smoothed the observations using the medians of time windows and removed the outliers; in the remote sensing data the texture features were processed to ensure they were orientation independent.

Feature selection is important in scientific applications for several reasons. Removing redundant or irrelevant features is likely to improve the accuracy of classifiers, which was the goal in FIRST and the human settlements applications. Identifying features crucial to classification can provide insights into the underlying phenomena, which is of interest to the scientists. In cases like the fusion problem, providing these insights is the goal of the project. Feature selection is also important because identifying which features are worth generating can save considerable resources. Our results with the remote sensing data, for example, indicate that we need to calculate features from only two out of four spectral bands and only the high-frequency components of the wavelet and Gabor features.

Our experience suggests that simple methods work well in many cases. While using non-linear classifiers and more sophisticated feature selection methods might result in higher classification accuracies, the results obtained with simpler techniques, such as the ones presented in this paper, are adequate to identify relevant features in many applications.

## Acknowledgments

We would like to thank the rest of the Sapphire team at LLNL (www.llnl.gov/casc/sapphire) for useful discussions and computational help. We gratefully acknowledge our FIRST collaborators R. Becker, M. Gregg, D. Helfand, S. Laurent-Muehleisen, and R. White for their technical interest and support of this work. We also thank K. H. Burrell and M. Walker for their help and support of the work with the fusion data. We acknowledge D. Poland for the tool used to generate the ground truth in the remote sensing application. Finally, we thank B. de Supinski for his comments on a draft of this manuscript.

UCRL-CONF-202657. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

## 7. REFERENCES

[1] R. H. Becker, R. White, and D. Helfand. The FIRST survey: Faint images of the radio sky at twenty-cm. *Astrophysical Journal*, 450:559–599, 1995.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. CRC Press, 1984.

[3] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In C. Brodley and A. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning*, pages 74–81, San Francisco, CA, 2001. Morgan Kaufmann Publishers.

[4] I. K. Fodor, E. Cantú-Paz, C. Kamath, and N. Tang. Finding bent-double radio galaxies: A case study in data mining. In *Interface: Computer Science and Statistics*, volume 33, 2000.

[5] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In L. Saitta, editor, *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, San Mateo, CA, 1996. Morgan Kaufmann.

[6] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:610–621, 1973.

[7] S. H. Huang. Dimensionality reduction on automatic knowledge acquisition: a simple greedy search approach. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1364–1373, 2003.

[8] S. Mallat. A theory for multi-resolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[9] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[10] K. Mardia, J. T. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1995.

[11] S. D. Newsam and C. Kamath. Retrieval using texture features in high resolution multi-spectral satellite imagery. In *SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI*, 2004.