CLASSIFYING BENT-DOUBLE GALAXIES

Astronomy data sets have led to interesting problems in mining scientific data. These problems will likely become more challenging as the astronomy community brings several surveys online as part of the National Virtual Observatory, giving rise to the possibility of mining data across many different surveys.

Data mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures in data. These techniques have long been applied to astronomy data, both by astronomers using data mining techniques and by data miners working with astronomy data. Such research has included the use of neural networks to discriminate between stars and galaxies, as well as the use of decision trees for star–galaxy classification, the identification of volcanoes in Venusian imagery, and the classification of galaxies with a bent-double morphology.[1–4] These efforts have typically focused on data in a single survey, although researchers often cross-verify their results with other surveys for validation.

The data in an astronomical survey is usually available in two forms: images and catalogs. The original data that the telescopes take (after some preprocessing) is in the form of images, which, taken together, tile a large area of the sky. Once the telescope obtains the images, the astronomer can create a catalog that includes information on each object in the image. Depending on the type of problem being solved through data mining, the catalog data could require further processing before it is ready for pattern identification. To successfully apply data mining to an astronomy problem, the data miner must understand the problem, the data, and the processing performed to generate the catalog from that data. This requires close collaboration with the astronomers in all aspects of the data mining process—from feature identification and extraction to result validation and process refinement.

In this article, we discuss the work we performed while using the catalog from the FIRST (Faint Images of the Radio Sky at Twenty centimeters) survey to classify galaxies with a bent-double morphology, meaning those galaxies that appear to be bent in shape. We describe the approach we took to mine this data, the issues we addressed in working with a real data set, and the lessons we learned in the process.

1521-9615/02/$17.00

CHANDRIKA KAMATH, ERICK CANTÚ-PAZ, IMOLA K. FODOR, AND NU AI TANG
Lawrence Livermore National Laboratory

FIRST data

The FIRST survey is a project that started in 1993 with the goal of producing the radio equivalent of the Palomar Observatory Sky Survey.[5] Using the National Radio Astronomy Ob-

servatory's very large array (VLA), FIRST is scheduled to cover more than 10,000 square degrees of the northern and southern galactic caps, to a flux density limit of 1.0 mJy (milli-Jansky). With the data from its first six years' worth of observations, FIRST has covered about 8,000 square degrees, producing more than 32,000 images, each with two million pixels. At a threshold of 1 mJy, approximately 90 radio-emitting galaxies, or radio sources, occupy a typical square degree. The results we present in this article are based on the 2000 version of the catalog, which includes data from 1993 through 1999. The survey is ongoing; new data is being collected, processed, and made available in the public domain at the FIRST Web site, http://sundog.stsci.edu.

Radio sources exhibit a wide range of morphological types that provide clues to the source class of the galaxy, its emission mechanism, and the properties of the surrounding medium. Radio sources with a bent-double morphology particularly interest astronomers because they indicate the presence of clusters of galaxies. FIRST scientists currently use a manual approach to detect bent-double galaxies. They look at a radio source's image, and if it appears to be a bent double, they cross-validate it with other surveys. This visual inspection of the radio images, besides being very subjective, is becoming increasingly infeasible as the survey grows. Our goal is to automate this process of classifying galaxies using techniques from data mining. We used as a training set the galaxies that astronomers had manually identified as bent doubles and nonbent doubles. We used this training set, along with the features representing each galaxy, to induce a decision tree model, which we then applied to the classification of unlabeled galaxies.

Figure 1 includes several examples of radio sources from the FIRST survey, including both bent doubles and nonbent doubles. Although some galaxies are relatively simple in shape, others can be rather complex.

Like other astronomy surveys, data from the FIRST survey is available in two forms: image maps and a catalog. A user-friendly Web interface enables easy access to radio sources at a given RA (right ascension, analogous to longitude) and Dec (declination, analogous to latitude) position in the sky. Figure 2 shows an image map containing examples of two bent doubles. These large image maps are mostly "empty"—that is, composed of background noise. Each map covers an area approximately 0.45 square degrees, with pixels that are 1.8 arcseconds wide. We get these image maps as a result of processing the raw data collected by the VLA telescopes.

In addition to image maps, the FIRST survey also provides a catalog,[6] which is created by FIRST astronomers by processing an image map to fit 2D elliptic Gaussian to each radio source. For example, the lower bent double in Figure 2 is approximated by more than seven Gaussians whereas the upper one is approximated by three Gaussians. Due to an upper limit on the number of Gaussians needed to fit each radio source, highly complex sources are not approximated well by using just the information in the catalog. Each entry in the catalog corresponds to the information on a single Gaussian. This includes, among other things, the RA and Dec for the Gaussian's center, the lengths of the major and minor axes, the peak flux, and the position angle of the major axis (degrees counterclockwise from north). A radio source is composed of one or more catalog entries. For the data collected through 1999, the image data is 250 Gbytes, whereas the catalog is much smaller—approximately 80 Mbytes.

Extracting features from the FIRST catalog

In our work on the bent-double problem, we decided that we would first focus on the catalog data. It was not only easier to work with the catalog, but the astronomers believed that it was a good approximation to all but the most complex radio sources. Therefore, our first task was to group the catalog entries—that is, the elliptic Gaussians—into radio sources. Our algorithm starts with an entry in the catalog, searches for other entries within a region of interest of 0.96 arc-minutes, restarts the search from each newly found entry, and repeats until it can't find any more new catalog entries within the region of interest. The algorithm collects all the catalog entries it finds in this search to form a radio source. Next, the algorithm repeats the entire grouping procedure, starting from the next available catalog entry and excluding any entries that are part of already existing radio sources.

Finding cases where two distinct galaxies are widely separated in the 3D sky but appear close to each other in the 2D projection of the FIRST images is not hard. For example, Figure 3, with the image centered at RA = $10^h50^m08.5^s$ and Dec = +30°40′15″ (Julian 2000 coordinates), shows

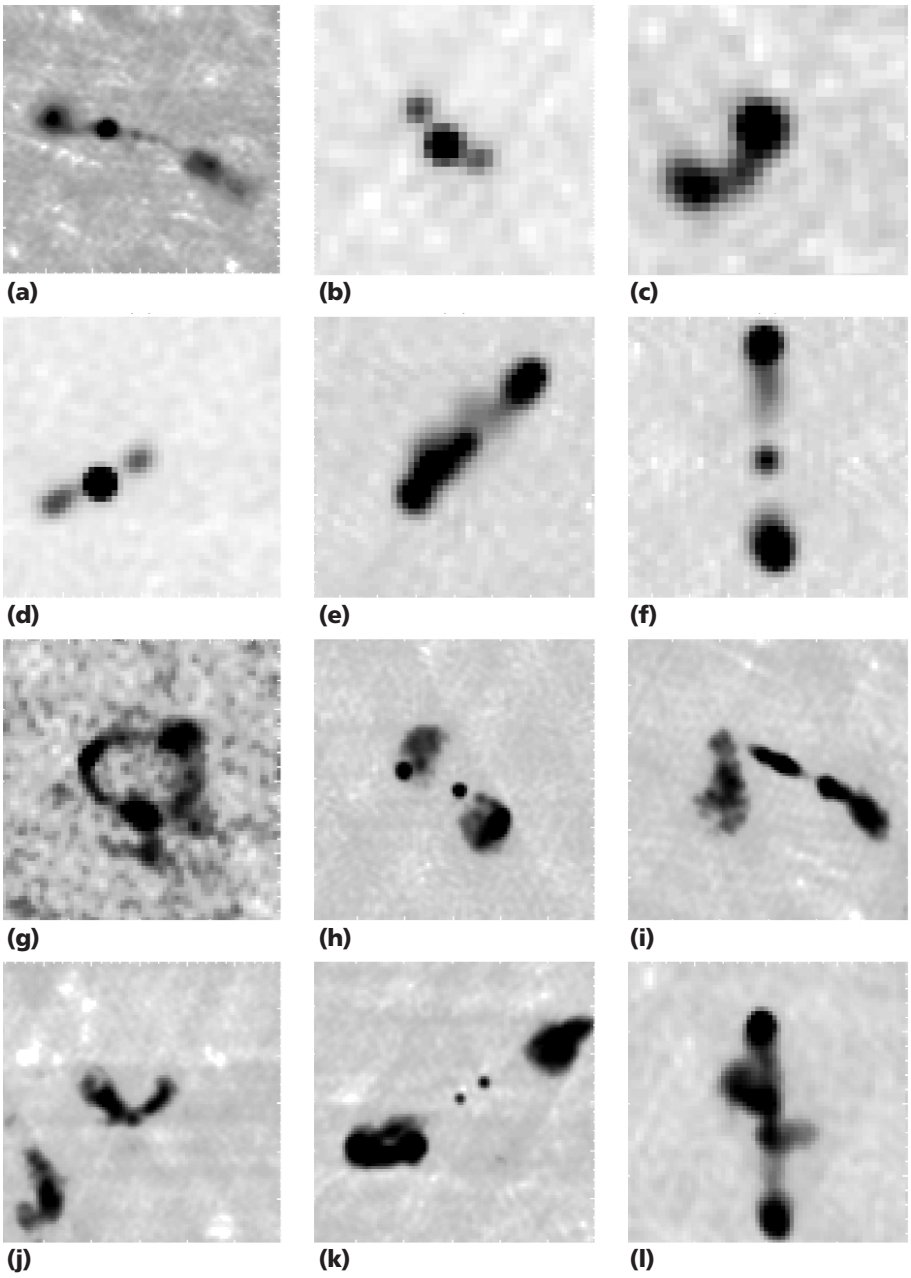
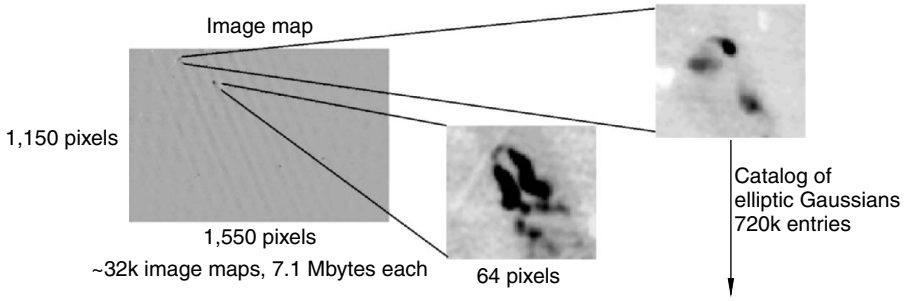


Figure 1. Example radio sources from the FIRST survey: (a) through (c) bent doubles; (d) through (f) nonbent doubles; and (g) through (l) complex sources. The similarity between a (b) bent double and a (d) nonbent double indicates one of the difficulties in automating the detection of bent-double galaxies.



Catalog entry	RA	Dec	Peak Flux (mJy/beam)	Major axis (arc-sec)	Minor axis (arc-sec)	Position angle (degrees)
Galaxy	00 56 25	–01 15 43	25.38	7.39	2.23	37.9
	00 56 26	–01 15 57	5.50	18.30	14.29	94.2
	00 56 24	–01 16 31	6.44	19.34	10.19	39.8

Figure 2. FIRST data: image maps and catalog entries.

two distinct radio sources within 0.96 arc-minutes of each other. Such examples illustrate why the task of automated bent-double detection is a rather hard problem, because processing techniques designed to work for most cases sometimes fail in others. It also shows the ease with which humans can visually identify the two objects as being separate, a task that is difficult to automate.

After grouping the catalog entries into complex radio sources, we separated the data depending on the number of catalog entries that make up the sources. We did this to reduce the number of galaxies we had to classify. First, we knew that by using features from only the catalog, there were unlikely to be any bent doubles in the single-catalog-entry sources. Second, there are relatively few three-plus-entry sources, all of which are "interesting" to the astronomers, regardless of whether they are bent doubles. Our software flagged and reported them to the FIRST astronomers. This approach also helped us address cases in which two radio sources are close to each other and each composed of at least two catalog entries. However, it did not address cases in which two disconnected, yet close, sources were approximated by two or three Gaussians. For the 2000 catalog, the number of radio sources as a function of the number of catalog entries comprising them is

# Catalog entries	# Radio sources
1	514637
2	66571
3	15059
3+	6333

Having removed the single-entry and the three-plus-entry radio sources from consideration, we further split the sources into two- and three-entry sources. We did this because the number of features extracted depends on the number of catalog entries, and we wanted a feature vector with a uniform length. However, it also meant that a small training set (313 examples) split further into smaller training sets of 118 examples for two-entry sources and 195 examples for three-entry sources. We initially focused on the three-entry sources because identifying features representing "bentness" seemed easier. We also had a larger training set, with 167 bent doubles and 28 nonbent doubles. Over 15,000 such three-entry galaxies appear in the data collected through 1999, making a visual inspection tedious. Moreover, the training set is unbalanced, with far more bent doubles than nonbent doubles.

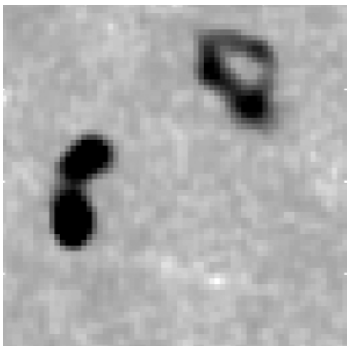


Figure 3. An example image from FIRST, illustrating two galaxies close together.

We identified the features for the bent-double problem through extensive conversations with FIRST astronomers. As we asked them to justify their decisions in identifying a radio source as a bent double, it became apparent that they placed great importance on spatial features such as distances and angles. Frequently, the astronomers would characterize a bent double as a radio-emitting "core," with one or more additional components at various angles that were usually side-wakes left by the core as it moved relative to the Earth. Once we extracted an initial set of features, we continued refining the features until the cross-validation error for a decision tree classifier reduced to about 10 percent, a number the astronomers felt was sufficient for their use.[7] We achieved the best accuracies by using just the features representing all three catalog entries simultaneously. Therefore, the remainder of this article focuses on these features.

In identifying the features based on all three catalog entries, we first identified the galaxy's core as the entry opposite the longest side of the triangle formed by the centers of the three Gaussians. Figure 4 depicts a possible arrangement of the three catalog entries, with entry A as the core.

Table 1 characterizes the features used.[8]

Decision trees

Let's examine our experimental results for the classification of bent-double galaxies with three catalog entries and various decision-tree based classifiers. These include a single tree and ensembles of 10 and 20 trees created using the Adaboost, bagging, and ArcX4 techniques for creating ensembles for classifiers.[9] We also developed an ensemble technique based on histograms.[10] In this method, instead of sorting each feature to determine the best split for it, we first create a histogram for each feature. Then, considering only

Table 1. Features used to represent a galaxy using all three catalog entries simultaneously.

Feature	Definition
totArea	The sum of the three areas of the elliptic Gaussians
peakFlux	The maximum of the three peak fluxes of the entries
sumIntFlux	The sum of the three integrated fluxes of the entries
avgDiffusion	The mean of the three diffusions, where $\text{diffusion}_X = \text{IntFlux}_X/\text{Area}_X$
totEllipt	The sum of the three ellipticities, where $\text{ellipticity}_X = \text{MajorAxis}_X/\text{MinorAxis}_X$
maxFlux	The maximum of the three integrated fluxes
coreAngl	The core angle, defined as angle BAC in Figure 4
angleAB	Angle ACB in Figure 4 (between sides *a* and *b*)
angleAC	Angle ABC in Figure 4 (between sides *a* and *c*)
totalBendGeom	The source's total bentness, equal to the sum of angles AMB and ANC
totalBendDiff	The source's total bentness, equal to the sum of |APosAngle – BPosAngle| and |APosAngle – CPosAngle|, where XPosAngle denotes the angle of the major axis of entry X, measured counterclockwise from north
ariAngl = acos BC/(AB + AC)	A measure of bentness[8]
ABAnglSide	The angle formed by the major axis of B with the AB segment, angle ABM
ACAnglSide	The angle formed by the major axis of C with the AC segment, angle ACN
sumComDist	The sum of the three pairwise distances between the centers of entries
sumRelDist	The sum of the three pairwise relative distances, calculated as $\frac{4XYComDist}{XMaj + XMin + YMaj + YMin}$ where for a pair of entries *X* and *Y*, *XYComDist* is the distance between their centers, and *XMaj*, *XMin* denote the major and minor axis of entry *X*
axialSym	A symmetry measure given by the ratio of the ellipticities of entries B and C
ariSym = AC/AB	A symmetry measure[8]
anotherSym = (AB + AC)/(AB + BC + AC)	Another symmetry measure
consDemote	{0/1} flag, 1 if one of the noncore entries is far from the core, and 0 otherwise (B is considered far if AB > 2 × const × (AMaj + BMaj), where const is currently set to 3 arc-seconds; similarly for C)

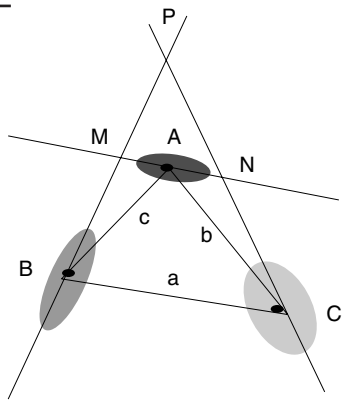


Figure 4. An example of the elliptic Gaussians fitted to a three-entry radio source.

the histogram bin boundaries as potential split points, we find the best such bin boundary. Next, to introduce randomness in the tree induction, we select an interval around the best bin boundary among all the features and select a point uniformly at random in this interval as the split point. In the experiments reported here, we used equal-width histograms, with the number of bins chosen to be the square root of the number of instances at a node. The interval's width is chosen to be the same as the bin's width. We refer to this technique as the *histogram-based ensemble*. In addition, we also consider the case of a single tree obtained by selecting the best bin boundary as the split point, without the randomization.

We first used our original set of 195 training examples to refine the set of features until the error rate dropped below 10 percent. Then, we used the tree created from this set of features to classify unlabeled galaxies. We showed several of these galaxies to the astronomers for validation. Because we wanted to use this new set of galaxies to enhance our training set, we selected a higher percentage of galaxies classified as nonbents. This process of validation is rather tedious and has the drawback of being subjective and somewhat inconsistent: the labels an astronomer as-

Table 2. Test error rates from different classification methods.

Method	Gini (%)	Gain ratio (%)	Information gain (%)
Single tree	32.41	42.76	30.00
Histogram-based tree	30.69	34.14	29.31
Histogram-based ensemble, 10 trees	33.17	27.97	32.93
	(0.597)	(0.854)	(0.565)
Histogram-based ensemble, 20 trees	32.24	26.55	32.28
	(0.465)	(0.465)	(0.507)
Adaboost, 10 trees	34.83	43.10	42.41
Adaboost, 20 trees	34.83	42.07	42.76
Bagging, 10 trees	34.72	33.79	36.65
	(0.974)	(1.087)	(1.61)
Bagging, 20 trees	32.38	32.24	33.65
	(0.486)	(0.478)	(0.681)
ArcX4, 10 trees	41.38	42.41	38.28
ArcX4, 20 trees	34.82	39.66	38.62

signs are subject to the drift common to human labelers. Therefore, we were able to validate only 290 galaxies, of which 92 were bents and 198 nonbents.

For the first experiment, we used these newly validated galaxies as a separate testing set. Table 2 gives the test error rates for this set of experiments; we got these results after 10 runs. In bagging and histogram-based ensemble techniques, the algorithm's randomization results in different test errors for each run. In these cases, the standard error is included as well. All results are presented for a tree or ensemble created without pruning; because the training set was rather small and unbalanced, pruning increased the error. We present results for three different splitting criteria: Gini, gain ratio, and information gain.[11]

The results in Table 2 indicate that if we create a classifier from the original 195 instances and test it on a different set of 290 instances, the error increases almost by a factor of 3. However, we should view these results with caution. First, the training set has far more bent doubles (167) than nonbent doubles (28). In contrast, the test set has more nonbents (198) than bents (92). As a result, we would expect poorer performance on the test set. Moreover, for some methods, the choice of a splitting criterion can affect the error rates substantially. The lowest error rate is for the histogram-based ensemble using the gain ratio criterion, whereas the other ensemble techniques fared rather poorly relative to a single tree.

To better compare the different methods, Table 3 lists a sample of the confusion matrices for each method, and Table 4 lists each method's precision, recall, and F-measure. For Table 3, the two columns of the top row list the number of bents identified as bent and nonbent, respectively. The two columns of the bottom row list the number of nonbents identified as bent and nonbent, respectively. The confusion matrix in Table 4 is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

where a and d are the correctly classified bents and nonbents, b is the number of bents classified as nonbent, and c is the number of nonbents classified as bents. Thus, precision, recall, and F-measure are

$$\text{Precision (P)} = a/(a + c)$$
$$\text{Recall (R)} = a/(a + b)$$
$$\text{F-measure} = 2PR/(P + R).$$

Here, we assume that the precision and recall are weighted equally. These results indicate that the histogram-based ensembles have the highest F-measure for the gain ratio splitting criterion, with bagging and the histogram-based single tree methods a close second. The relatively low precision indicates several false positives, but the higher recall indicates that few positives (bent doubles) are missed.

For our second experiment, we combined the newly validated set of 290 examples with the original training set of 195 instances and used the combined set of 485 instances to evaluate the algorithms with cross-validation. Because the data set is now larger, we include results with pessimistic error pruning because it improves

Table 3. Typical confusion matrices (bent, nonbent) for the different classification methods.

Method	Gini	Gain ratio	Information gain
Single tree	[74 18; 76 122]	[82 10; 114 84]	[72 20; 67 131]
Histogram-based tree	[71 21; 68 130]	[80 12; 87 111]	[71 21; 64 134]
Histogram-based ensemble, 10 trees	[65 27; 76 122]	[74 18; 53 145]	[74 18; 72 126]
Histogram-based ensemble, 20 trees	[72 20; 70 128]	[77 15; 60 138]	[71 21; 69 129]
Adaboost, 10 trees	[80 12; 89 109]	[81 11; 114 84]	[78 14; 109 89]
Adaboost, 20 trees	[80 12; 89 109]	[80 12; 110 88]	[77 15; 109 89]
Bagging, 10 trees	[77 15; 82 116]	[75 17; 63 135]	[72 20; 83 115]
Bagging, 20 trees	[75 17; 75 123]	[74 18; 76 122]	[78 14; 89 109]
ArcX4, 10 trees	[80 12; 108 90]	[79 13; 110 88]	[80 12; 99 99]
ArcX4, 20 trees	[81 11; 90 108]	[79 13; 102 96]	[79 13; 99 99]

performance. We also restricted our study to ensembles of 10 trees.

From the results presented in Table 5, we make the following observations. For this data set, with the given set of features and training examples, the error rates range from a low of 18.02 to a high of 22.79. The values in bold indicate the best error rate for each method across

Table 4. The precision, recall, and F-measure corresponding to Table 3's confusion matrices.

Method	Gini			Gain ratio			Information gain		
	P	R	F-measure	P	R	F-measure	P	R	F-measure
Single tree	0.49	0.80	0.61	0.41	0.89	0.56	0.52	0.78	0.62
Histogram-based tree	0.51	0.77	0.61	0.48	0.87	0.62	0.53	0.77	0.63
Histogram-based ensemble, 10 trees	0.46	0.71	0.55	0.58	0.80	0.67	0.54	0.80	0.64
Histogram-based ensemble, 20 trees	0.51	0.78	0.62	0.56	0.84	0.67	0.51	0.77	0.61
Adaboost, 10 trees	0.47	0.87	0.61	0.41	0.88	0.56	0.42	0.85	0.56
Adaboost, 20 trees	0.47	0.87	0.61	0.42	0.87	0.57	0.41	0.84	0.55
Bagging, 10 trees	0.48	0.84	0.61	0.54	0.81	0.65	0.46	0.78	0.58
Bagging, 20 trees	0.50	0.81	0.62	0.49	0.80	0.61	0.47	0.85	0.60
ArcX4, 10 trees	0.42	0.87	0.57	0.42	0.86	0.56	0.45	0.87	0.59
ArcX4, 20 trees	0.47	0.88	0.61	0.44	0.86	0.58	0.44	0.86	0.58

Table 5. Cross-validation error rates using different classification methods.

Method	Gini		Gain ratio		Information gain	
	No pruning	Pruning	No pruning	Pruning	No pruning	Pruning
Single tree	22.79	19.77	22.62	19.83	22.77	**19.71**
	(0.31)	(0.18)	(0.27)	(0.15)	(0.39)	(0.41)
Histogram-based, single tree	21.73	20.46	20.85	**18.81**	22.56	20.96
	(0.34)	(0.29)	(0.39)	(0.17)	(0.32)	(0.39)
Histogram-based, 10 trees	18.69	18.27	18.10	**18.02**	18.22	18.42
	(0.28)	(0.30)	(0.16)	(0.22)	(0.18)	(0.34)
Adaboost, 10 trees	21.87	**20.40**	22.37	22.56	20.50	20.75
	(0.42)	(0.45)	(0.53)	(0.47)	(0.45)	(0.43)
Bagging, 10 trees	19.40	18.35	18.98	**18.12**	18.98	18.52
	(0.28)	(0.34)	(0.34)	(0.26)	(0.36)	(0.35)
ArcX4, 10 trees	20.48	20.12	21.67	22.48	21.06	**19.77**
	(0.39)	(0.20)	(0.35)	(0.41)	(0.22)	(0.37)

the different splitting criteria and pruning options. When we consider the best error rate for each method across the different splitting criteria and pruning options, the histogram-based ensemble (18.02) and bagging (18.12) methods are the best performers. They improved the performance over the best single tree (19.71). In fact, even a single histogram-based tree (18.81) improved over a single tree. In comparison, Adaboost and ArcX4 did not perform as well.

Our experiences with mining the FIRST data to classify galaxies with a bent-double morphology led to several interesting observations.

First, astronomy data is frequently available in the form of images or catalog data from which all the relevant features have not been extracted. We found that identifying and extracting such features in a robust manner is nontrivial and results in one of data mining's more time-consuming steps. In our specific example of the detection of bent doubles, having the catalog immensely helped us get a head start on the problem because we did not have to work with the images themselves.

Second, we found that FIRST data's availability on the Web, as well as tools to read, write, and display it were very helpful. This is not always the case with many data sets in astronomy and other sciences.

Third, it took us almost six months to understand the problem domain, the problem, and the data itself. During this time, we had extensive conversations with the FIRST astronomers. This understanding of the data and close collaboration with the domain scientists is an important but often overlooked aspect of data mining, especially in scientific domains.

Fourth, in our classification problem, we found a dearth of labeled examples because the astronomers had to manually identify them. Furthermore, the labels for the galaxies were subjective, and we found astronomers sometimes disagreed on whether to classify a galaxy as a bent double or a nonbent double. This was especially true in difficult-to-classify cases. Given the lack of ground truth in this problem, generating a good training set was difficult. This sharply contrasts with commercial data, where labeled examples might have been generated historically—for example, in customer churn problems.

Finally, because FIRST is a survey in progress, we found that our bookkeeping had to keep up with changes made in different data releases. For example, the 1999 version of the data merged information from the northern and southern hemispheres that was previously separate. Also, in the 2000 version, we found that certain galaxies no longer appeared in the catalog. Conversations with astronomers indicated that this was a normal occurrence as a result of the processing of the data the telescopes collected. For galaxies at the very edge of the survey one year, additional data collected the following year made the pixels corresponding to those galaxies fall below the detection threshold. This meant that we had to be careful in our use of the ID tags for galaxies as we moved to newer versions of the survey.

Our next step in this work is to apply these decision tree techniques to the case of two-catalog-entry galaxies. We expect the problem to be

harder because the catalog might not capture all the features needed to discriminate between bent doubles and nonbent doubles. We are also applying these techniques to data from remote sensing and simulations of turbulent flow.

Acknowledgments

We gratefully thank our FIRST collaborators Robert Becker, Michael Gregg, David Helfand, Sally Laurent-Muehleisen, and Richard White for their technical interest and support. We also thank Charles Musick, Deanne Proctor, and Ari Buchalter for useful discussions and computational help.

We performed this work under the auspices of the US Department of Energy, University of California Lawrence Livermore National Laboratory, under contract W-7405-Eng-48, UCRL-JC-147497.

References

1. S. Odewahn et al., "Automated Star/Galaxy Discrimination with Neural Networks," *Astronomical J.*, vol. 103, no. 1, 1992, pp. 318–331.
2. N. Weir, U. Fayyad, and S. Djorgovski, "Automated Star/Galaxy Classification for Digitized DPOSS-II," *Astronomical J.*, vol. 109, no. 6, 1995, pp. 2401–2414.
3. M. Burl et al., "Learning to Recognize Volcanoes on Venus," *Machine Learning*, vol. 30, nos. 2–3, 1998, pp. 165–195.
4. C. Kamath et al., "Searching for Bent-Double Galaxies in the FIRST Survey," *Data Mining for Scientific and Eng. Applications*, R. Grossman et al., eds., Kluwer, Boston, 2001, pp. 95–114.
5. R.H. Becker, R. White, and D. Helfand, "The FIRST Survey: Faint Images of the Radio Sky at Twenty-cm," *Astrophysical J.*, vol. 450, 1995, pp. 559–577.
6. R.L. White et al., "A Catalog of 1.4 GHz Radio Sources from the FIRST Survey," *Astrophysical J.*, vol. 475, 1997, pp. 479–493.
7. I.K. Fodor et al., "Finding Bent-Double Radio Galaxies: A Case Study in Data Mining," *Computing Science and Statistics*, vol. 32, 2000, pp. 37–47.
8. J. Lehár et al., "An Efficient Search for Gravitationally Lensed Radio Lobes," *Astrophysical J.*, vol. 547, 2001, pp. 60–76.
9. D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *J. Artificial Intelligence Research*, vol. 11, 1999, pp. 169–198.
10. C. Kamath, E. Cantú-Paz, and D. Littau, "Approximate Splitting for Ensembles of Trees Using Histograms," *Proc. 2nd SIAM Int'l Conf. Data Mining*, 2002, pp. 370–383.
11. L. Breiman et al., *Classification and Regression Trees*, CRC Press, Boca Raton, Fla., 1984.

Chandrika Kamath is a computer scientist and project lead at the Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. Her research interests include image processing, pattern recognition, and practical applications of data mining. She received a PhD in computer science from the University of Illinois, Urbana-Champaign. She is a member of the ACM, the IEEE, SIAM, and SPIE. Contact her at the Center for Applied Scientific Computing, LLNL, 7000 East Ave., PO Box 808, L-561, Livermore, CA 94551; kamath2@llnl.gov.

Erick Cantú-Paz is a computer scientist at the Center for Applied Scientific Computing, LLNL. His research interests include theoretical foundations and practical applications of evolutionary computation, machine learning, and data mining. He received a PhD in computer science from the University of Illinois, Urbana-Champaign. He is a member of the ACM, the IEEE, and the International Society of Genetic and Evolutionary Computation. Contact him at the Center for Applied Scientific Computing, LLNL, 7000 East Ave., PO Box 808, L-561, Livermore, CA 94551; cantupaz@llnl.gov.

Imola K. Fodor is a computational mathematician at the Center for Applied Scientific Computing, LLNL. Her research interests include signal processing and statistical issues in data mining. She received a PhD in statistics from the University of California, Berkeley. She is member of the Society for Industrial and Applied Mathematics, the American Statistical Association, and the Institute of Mathematical Statistics. Contact her at the Center for Applied Scientific Computing, LLNL, 7000 East Ave., PO Box 808, L-560, Livermore CA 94551; fodor1@llnl.gov.

Nu Ai Tang is a computer scientist at the Center for Applied Scientific Computing, LLNL. Her research interests include object-oriented programming and data mining. She received a BS in computer science from California State University, Sacramento. She is a member of the Society of Women Engineers. Contact her at the Center for Applied Scientific Computing, LLNL, 7000 East Ave., PO Box 808, L-560, Livermore CA 94551; tangn@llnl.gov.