Finding Bent-Double Radio Galaxies: A Case Study in Data Mining *

I. K. Fodor, E. Cantú-Paz, C. Kamath, N. A. Tang
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA 94551

Abstract

This paper presents our early results in applying data mining techniques to the problem of finding radio-emitting galaxies with a bent-double morphology. In the past, astronomers on the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey have detected such galaxies by first inspecting the radio images visually to identify probable bent-doubles, and then conducting observations to confirm that the galaxy is indeed a bent-double. Our goal is to replace this visual inspection by a semi-automated approach. In this paper, we present a brief overview of data mining, describe the features we use to discriminate bent-doubles from non-bent-doubles, and discuss the challenges faced in defining meaningful features in a robust manner. Our experiments show that data mining, using decision trees, can indeed be a viable alternative to the visual identification of bent-double galaxies.

1 Introduction

Data mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures and events in data (Kamath and Musick, 2000, and the references therein). One of the steps in data mining is pattern recognition, where a pattern is identified using measurable features or attributes extracted from the data. Data mining, as illustrated in Figure 1, is an interactive and iterative process involving data preprocessing, search for patterns, and interpretation of the results. Input from domain scientists is an integral part of the data mining process, and frequently results in the refinement of one or more steps. The preprocessing of the data is a very important and time consuming first step as features relevant to the pattern have to be extracted from raw input data.

As part of the Sapphire project at LLNL (http://www.llnl.gov/casc/sapphire), we are developing an object-oriented parallel framework for mining massive scientific data sets. One of these data sets is the FIRST survey, where we are interested in identifying radio-emitting galaxies with a bent-double morphology. In this paper we describe our early experiences in applying data mining techniques to solve this problem. We show that the success of a pattern recognition technique, such as decision trees, is dependent on the features we have extracted from the raw data. Finding relevant features that are scale, rotation, and translation invariant is non-trivial. Further, defining these features in a robust and consistent manner can be a challenge as well.

The outline of this paper is as follows: Section 2 describes the FIRST survey, and outlines the problem of detecting bent-double radio galaxies. Section 3 provides details on the approach we have taken to address the difficulties encountered in solving this problem. Section 4 reports our results, focusing on the important role played by the data preprocessing step in data mining. Section 5 concludes with a summary and future work.

2 The FIRST Survey

The FIRST — Faint Images of the Radio Sky at Twenty-cm — survey (Becker et al. 1995) is a project that was started in 1993 with the goal of producing the radio equivalent of the Palomar Observatory Sky Survey. Using the Very Large Array (VLA) at the National Radio Astronomy Observatory (NRAO), 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). At present, with the data from the 1993 through 1998 observations, FIRST has covered about 6,000 square degrees, producing more than 20,000 two-million pixel images. At a threshold of 1mJy, there are approximately 90 radio emitting galaxies, or radio sources, in a typical square degree. Note that the results of this paper are based on the 1998 catalog, including data from 1993-1997.

*Accepted for publication in Computing Science and Statistics, Volume 33, 2000

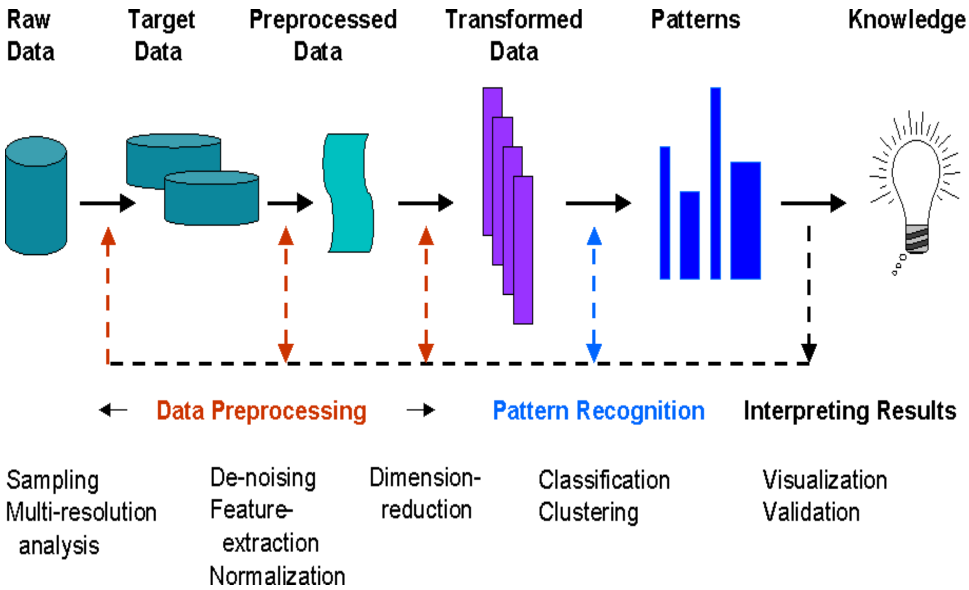
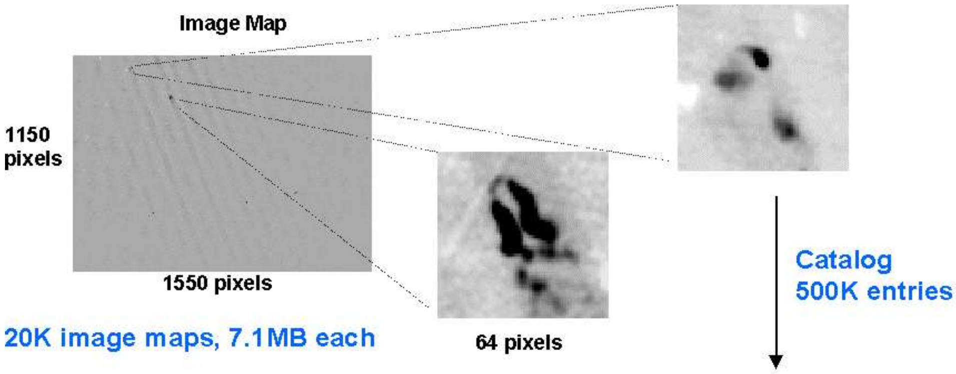


Figure 1: Data mining: an iterative and interactive process.



RA	DEC	Peak Flux (mJy/bm)	Major Axis (arcsec)	Minor Axis (arcsec)	Position Angle (degrees)
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: Examples of FIRST images and catalog entries.

Radio sources exhibit a wide range of morphological types that provide clues to the source class, emission mechanism, and properties of the surrounding medium. Of particular interest are sources with a bent-double morphology as they indicate the presence of clusters of galaxies, a key project within the FIRST survey. The current approach used by FIRST scientists for the detection of bent-doubles is a manual one. They first look at the image of a radio source to see if it could be labeled as a bent-double. If two out of three astronomers agree that the galaxy is a bent-double, then additional observations are carried out in order to study the bent-double in more detail.

The visual inspection of the radio images, besides being very subjective, is also becoming increasingly infeasible as the survey grows in size. Our long-term goal is to automate this process of classifying galaxies as bent-doubles using techniques from data mining.

2.1 Data from the FIRST Survey

The data from FIRST, both raw and postprocessed, are readily available on the FIRST website (http://sundog.stsci.edu/). A user friendly 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.

There are two forms of data available for use — image maps and a catalog. For example, in Figure 2, we show 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 approximately 0.45 square degrees area of the sky, and has pixels which are 1.8 arc seconds wide.

In addition to the image maps, FIRST also provides a source catalog (White et al. 1997). This catalog is obtained by processing an image map in order to fit two-dimensional elliptic Gaussians to each radio source. For example, the lower bent-double in Figure 2 is approximated by more than seven Gaussians while the upper one is approximated by three Gaussians. There is an upper limit to the number of Gaussians that are used to fit each radio source. As a result, highly complex sources are not approximated well 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 center of the Gaussian, the major and minor axes, the peak flux, and the position angle of the major axis (degrees counterclockwise from North).

3 Identifying Bent-Doubles

As illustrated in Figure 2, we have data at two extremes: image maps totaling about 200 Gigabytes, with the very few "interesting" pixels corresponding to the radio sources, and the 59 Megabyte catalog data with information about parts of a radio source. We could use either, or both, of these data to extract the features for the identification of the bent doubles. We decided to start with the features from the catalog for several reasons:

- The astronomers believed that the catalog was a good approximation to all but the most complex of radio sources.
- It was easier to work with the catalog as it was smaller.
- Processing the very large image maps for extracting relevant features for the bent-double problem was expected to be difficult and time consuming due to lack of parallel image processing software.
- The FIRST astronomers indicated that several of the features they thought were important in identifying bent-doubles were easily calculated from the catalog.

In this paper, we present the results using features based only on the catalog. In Section 4, we comment on the effects of this decision.

Having decided to work with the information in the FIRST catalog, the first step in classifying the bent-doubles was to group the catalog entries, i.e. 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 no more catalog entries are found within the region of interest. All catalog entries found in this search are collected to form a radio source. Next, the algorithm repeats the entire grouping procedure starting from the next available catalog entry, excluding the entries that are part of already existing radio sources.

In grouping the entries, once a new entry was found within the region of interest, the search could continue from either 1) the center of the new entry, or, 2) the center of mass of the entries that make up the source. Our experience indicated that the choice of the starting point had little effect on the resulting grouping.

Note that it is not very difficult to find cases where the catalog entries from one radio source are within 0.96 arc minutes of the catalog entries of a different radio source.

For example, Figure 3 with the image centered at RA $= 10^{h}50^{m}08.5^{s}$ and Dec $= +30^{o}40'15''$ (J2000 coordinates), shows two radio sources, a bent-double in the lower left corner, and a ring-like structure in the upper right corner. While in three dimensions, these radio sources may be far from each other, in a two-dimensional projection, they appear close together. Such examples illustrate why the task of automated detection of bent doubles is a rather hard problem, and one reason why visual inspection of an image is followed by detailed observations to confirm that the galaxy is a bent-double.

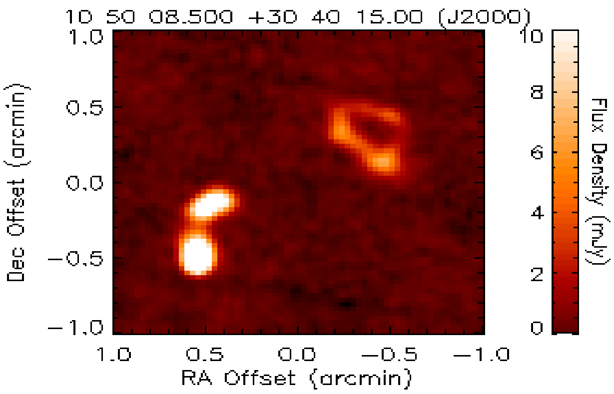


Figure 3: An example image from the FIRST website.

After grouping the entries into complex radio sources, we separated the data depending on the number of catalog entries making up the sources. There is a data set each for all the 1-entry sources, all the 2-entry sources, all the 3-entry sources, and all the 4plus-entry sources. This separation by the number of catalog entries was done for several reasons. First, we knew that, using features from only the catalog, there were unlikely to be any "bent-doubles" in the single catalog entry sources. This was because a single elliptic Gaussian could not be "bent". Further, there are relatively few 4plus-entry sources, all of which are "interesting" to the astronomers, regardless of whether they are bent-doubles or not. So, we simply flag them and report them to the scientists. This approach also helped us to address the case where there are two radio sources close to each other, with each composed of at least two catalog entries.

Having removed the single-entry, and the 4plus-entry radio sources from consideration, we further split the sources into two- and three-entry sources. This was done as the number of features extracted depends on the number of catalog entries, and we wanted a feature vector with a uniform length. However, this also meant that the size of the training set for the detection of bent-doubles was now divided into smaller training sets.

For the 1998 catalog, including observations from 1993 through 1997, the number of radio sources as a function of the number of catalog entries they are composed of, is as follows:

# Catalog entries	# Radio sources
1	311785
2	40134
3	9235
4+	4765

Once the radio sources (including the training set) were separated based on the number of catalog entries in the galaxy, we derived the features listed in Section 3.1 for the two and three entry sources. Next, using the appropriate training set, we created the decision trees for the identification of two and three entry radio sources. These trees were created using the C5.0 decision tree software package (Rulequest Research, http://www.rulequest.com). We also ran cross-validation experiments to determine the accuracy of the tree as the features used were varied.

We are currently in the process of running the most accurate decision tree, constructed with C5.0 from the initial training set, on the unlabeled radio sources. We plan to show a small sample of the new bent and non-bent-doubles to the astronomers and use their input to enhance the training set. The process will then be repeated until we have a large enough training set to have confidence in the tree generated.

3.1 Features for Bent-Doubles

This section describes various potential features that might be used to discriminate galaxies with bent-double morphology. Some of the features are directly taken from the FIRST catalog, some are derived from the basic ones in the catalog, and some are closely related. Note that we keep a few "features" for bookkeeping purposes only. Our focus is on features that are scale, rotation and translation invariant, as the pattern we are looking for, namely the bent-double, is scale, rotation, and translation invariant. We are also interested in features that are robust, that is, not sensitive to small changes in the data (White, 1999). Of course, it goes without saying that the features we select must be relevant to the problem.

We identified the features for the bent-double problem through extensive conversations with FIRST astronomers. As we asked them to justify their decision in identifying a radio source as a bent-double, it became apparent that greater focus was placed 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, which were usually side-wakes left by the core as it moved relative to the Earth.

We next list all the features we calculated based on our conversations with the astronomers. These features have been included here both for illustrative purposes, and for future reference. However, as some of them do not scale, or are not rotation invariant, it does not make sense to include them all in constructing the decision tree.

3.1.1 Features per Catalog Entry

The following list enumerates potential features pertaining to a single catalog entry.

1. **peak_flux**: the peak flux value (mJy)
2. **total_area** = $\frac{\pi \textbf{ maj min}}{4}$: the total area of the entry, as measured by the fitted elliptical Gaussian, where **maj** and **min** are the lengths of the major and of the minor axes, respectively
3. **int_flux**: the integrated flux value (mJy)
4. **ra**: the right ascension RA (decimal hours)
5. **dec**: the declination Dec (decimal hours)
6. **ellipticity** = $\frac{\textbf{maj}}{\textbf{min}} \geq 1$: a measure of the the entry's ellipticity, with one being a circular entry
7. **rms**: the local noise estimate (mJy) at the position of the entry in the sky
8. **sidelobe**: $\{0/1\}$ flag, 1 if the entry might be a sidelobe of a nearby bright source, 0 otherwise
9. **maj**: the size of the major axis (arc seconds)
10. **min**: the size of the minor axis (arc seconds)
11. **diffusion** = $\frac{\textbf{int\_flux}}{\textbf{total\_area}}$: a measure of diffusion
12. **SNR** = $\frac{\textbf{peak\_flux}-0.25}{\textbf{rms}}$: the peak flux density signal to noise ratio (the 0.25 reflects a bias correction documented in the FIRST survey); it can also be thought of as a "standardized" peak flux quantity
13. **point_source**: $\{0/1\}$ flag, 1 if the entry is a point source (its **maj** less than 2 arc seconds), and 0 otherwise
14. **flux**: set to **peak_flux** for point sources, and to **int_flux** for extended sources

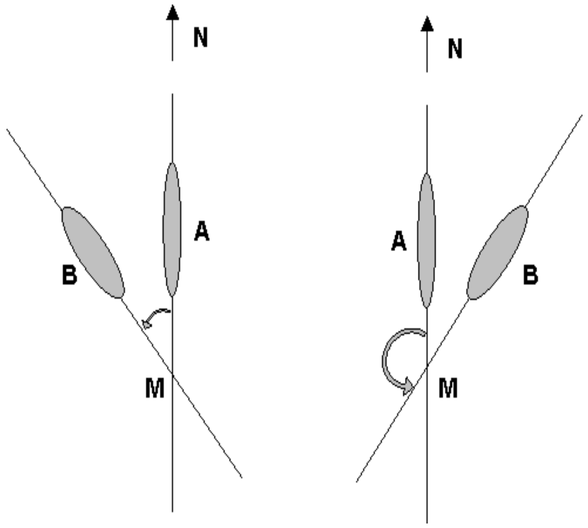


Figure 4: Two examples of 2-entry fitted radio sources.

15. **position_angle**: the angle (degrees) of the major axis, measured counterclockwise from North — for entry B, the angles indicated by an arrow in Figure 4, about 45^o in the left, and about $(180-45)^o$ in the right (0 for entry A in both cases)

Note that **maj** and **min** denote the entire lengths of the axes, i.e. the lengths between the two intersection points of the corresponding axis and the ellipse (as opposed to the lengths from the center of the ellipse to one of the intersection points). Also, the detection limit of the survey is about 2 arc seconds, so any **maj** or **min** less than 2 arc seconds is set to 2 arc seconds.

3.1.2 Pairwise Features

The potential features for a 2-entry radio source or two catalog entries as a pair are listed below. Features 1 through 9 characterize a 2-entry radio source, and features 10 through 22 pertain to any two entries taken together. This distinction will become clearer in the 3-entry radio source case, Section 3.1.3, where the meaning of features 4-9 will change to include all three components, and the features will include all three combinations of the last 13 pairwise features. Figure 4 shows two possible geometries for fitted 2-entry sources (i.e. the fitted elliptical Gaussians in the plane).

1. **id**: radio source identification number, for bookkeeping purposes only
2. **hemisphere**: radio source hemisphere, for bookkeeping purposes only

3. **num_ce**=2: number of entries in the source, for bookkeeping purposes only

4. **total_area**: the sum of the two total areas

5. **peak_flux**: the max of the two peak fluxes

6. **sum_int_flux**: the sum of the two integrated fluxes

7. **avg_diffusion**: the mean of the two diffusions

8. **tot_elliptic**: the sum of the two ellipticities

9. **flux**: the max of the two fluxes

10. **com_dist**: distance between the two centers

11. **rel_dist** $= \frac{4\ \mathbf{com\_dist_{12}}}{\mathbf{maj_1+min_1+maj_2+min_2}}$: a measure of the relative distance between the two entries, values close to one indicating nearly intersecting entries

12. **rel_pflux**: ratio of the two peak fluxes

13. **rel_flux**: ratio of the two fluxes

14. **rel_maj**: ratio of the two majors

15. **rel_iflux**: ratio of the two integrated fluxes

16. **rel_ellip**: ratio of the two ellipticities

17. **pair_angle_geom**: angle formed by the position angles of the two major axes, as calculated geometrically – angle AMB in both panels of Figure 4

18. **pair_angle_diff**: angle formed by the position angles of the two major axes, as calculated by the absolute difference in the two position angles — about $|0-45|^o = 45^o$ in the left, and about $|0-135|^o = 135^o$ in the right panel of Figure 4

19. **angle_flag**: $\{0/1\}$ flag, 1 if **pair_angle_geom** is unstable (i.e., could flip from α to $180-\alpha$), and 0 otherwise

20. **avg_SNR**: the mean of the two signal to noise ratios

21. **max_SNR**: the largest of the two signal to noise ratios

22. **rel_SNR**: the ratio of the two signal to noise ratios

The features for 2-entry radio sources include the 22 pairwise features above and the 2×15 single entry features for the two components listed in Section 3.1.1. The ordering in 2-entry sources is according to the maximum integrated flux. For 3-entry sources, the pairwise features are given in certain orders, depending on the ordering scheme - see Section 3.1.3.

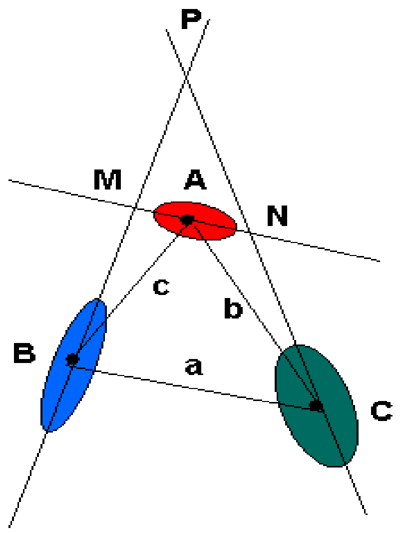


Figure 5: An example of a 3-entry fitted radio source.

3.1.3 Triple Features

Potential features characterizing three catalog entries are reported below. These descriptions assume the entries are ordered such that entry A is the core, and ABC denotes the triangle formed by the three entry centers. The two non-core entries, B and C, can be ordered in different ways, as explained in the paragraph following the list. Figure 5 depicts a characteristic fitted 3-entry source geometry. The notation of the following features refer to Figure 5.

1. **id**: radio source id, for bookkeeping purposes only

2. **hemisphere**: radio source hemisphere for bookkeeping purposes only

3. **num_ce**=3: number of entries in the source, for bookkeeping purposes only

4. **total_area**: the sum of the three total areas

5. **peak_flux**: the max of the three peak fluxes

6. **sum_int_flux**: the sum of the three integrated fluxes

7. **avg_diffusion**: the mean of the three diffusions

8. **tot_elliptic**: the sum of the three ellipticities

9. **flux**: the max of the three fluxes

10. **core_angle**: the core angle, defined as the angle BAC in the triangle above

11. **angle_ab**: angle ACB in the triangle above (between sides a and b)

12. **angle_ac**: angle ABC in the triangle above (between sides a and c)

13. **total_bend_geom**: the total bentness of the source, as measured by the sum of the two pairwise angles $\mathbf{pair\_angle\_geom}_{A,B} = AMB$ and $\mathbf{pair\_angle\_geom}_{A,C} = ANC$

14. **total_bend_diff**: the total bentness of the source, as measured by the sum of the two pairwise angles $\mathbf{pair\_angle\_diff}_{A,B} = |\mathbf{position\_angle_A} - \mathbf{position\_angle_B}|$ and $\mathbf{pair\_angle\_diff}_{A,C} = |\mathbf{position\_angle_A} - \mathbf{position\_angle_C}|$

15. $\mathbf{ari\_angle} = \mathrm{acos}\frac{\mathrm{BC}}{\mathrm{AB+AC}}$: a measure of bentness (Lehar et al. 2000) due to Ari Buchalter (originally proposed if $\mathrm{BC} \geq \max\{\mathrm{AB}, \mathrm{AC}\}$, i.e., the entries are ordered as in II. in the next paragraph)

16. **sum_com_dist**: the sum of the three pairwise **com_dist**

17. **sum_rel_dist**: the sum of the three pairwise **rel_dist**

18. **axial_sym**: a symmetry measure given by the ratio of the ellipticities of entries B and C

19. $\mathbf{ari\_sym} = \frac{\mathrm{AC}}{\mathrm{AB}}$: a symmetry measure (Lehar et al. 2000) due to Ari Buchalter (originally proposed if $\mathrm{BC} \geq \mathrm{AB} \geq \mathrm{AC}$, i.e., the entries are ordered as in II. in the next paragraph)

20. $\mathbf{another\_sym} = \frac{\mathrm{AB+AC}}{\mathrm{AB+BC+AC}}$: another symmetry measure

21. **cons_demote**: $\{0/1\}$ flag, 1 if one of the non-core entries is far from the core, and 0 otherwise [B is considered far if $\mathrm{AB} > 2 \times \mathrm{const} \times (\mathbf{maj}_\mathrm{A} + \mathbf{maj}_\mathrm{B})$, where const is currently set to 3 arc seconds, and similarly for C]

The features for 3-entry radio sources include the 21 triple features above, the $3 \times$ (last 13) pairwise features listed in Section 3.1.2, and the 3×15 single features listed in Section 3.1.1.

There are various ways of selecting the core and ordering the entries in 3-entry radio sources. We considered the three methods described below.

I. Choose the entry with the largest integrated flux as the core. Order the entries as: A (maximum integrated flux), B (second largest integrated flux), C (smallest integrated flux). Note that this is a somewhat ad-hoc ordering, with no real astronomical basis behind it.

II. Choose the core as the entry opposite the largest side of the triangle formed by the centers of the three ellipses. Order the entries as: A (opposite largest side), B (opposite second largest side), C (opposite smallest side).

III. Choose the core to be the entry opposite the side that is most unlike the other two sides. Order the entries as: A (the center such that the two sides of the triangle that meet at that center are closest in length), B (the center such that the two sides of the triangle that meet at that center are second closest in length), C (the center such that the two sides of the triangle that meet at that center are farthest in length).

For the 3-entry sources, we repeated the feature extraction step separately for each of the three ordering methods, and ran the decision tree algorithm on the three different sets of features.

4 Results Using Decision Trees

As mentioned earlier, we expect that some of the features will not be important in finding bent-doubles. For example, the position in the sky, that is, the (RA, Dec) coordinates, should not influence the results, at least as long as bent-doubles are approximately randomly distributed over the celestial sphere. However, our initial experiments with decision trees indicated that the coordinates were influential. On further investigation, we realized that when the astronomers provided us examples of non-bent-doubles to use in our training set, they had focused on a small section of the sky, thus making the coordinates influential. In this case, the decision tree was "right", but there was a problem in the features we used in training. While we expected the decision tree to focus on the features which are discriminating, this experiment illustrated the important role played by domain knowledge in the selection of features. As a result, in the remaining sections, we exclude all the bookkeeping "features" and the position coordinates from the analyses.

We next summarize the preliminary results of our experiments on the bent-double problem. We first make the following observations:

- We are working with a relatively small training set (118 examples for two-catalog entry sources, and

195 for the three-catalog entry sources). As the bent and non-bent-doubles have to be manually labeled by FIRST scientists, putting together an adequate training set is a non-trivial task. As explained earlier, we plan to enhance our small training set by using feedback from the astronomers on the results of the preliminary decision trees.

- Scientists are often subjective in their labeling of galaxies as bent or non-bent. This would imply that the training set itself is not very accurate.
- We are currently using features from only the catalog. We would therefore expect that if the "bentness" of a radio-source was adequately captured by the catalog, we would do well in identifying a bent-double.

Our initial experiments found that these observations played an important role in the case of the two-entry radio sources. Using only catalog-based features with the limited training set, the decision trees created were erratic. In cross validation experiments, we found that the tree strongly depended on the subset selected from the full training set. The misclassification errors that resulted were also relatively high, on the order of 20%. We therefore defer a detailed analysis of the 2-entry sources until later, when we can refine the features, add image-based features, and increase the training sample.

We next present the results for the 3-entry radio sources.

4.1 3-Entry Sources

For the three catalog entry sources, the training set consists of 195 labeled examples, with 167 bent-doubles and 28 non-bent-doubles.

Using the features and methods listed in Section 3.1.3, we repeated 10-fold cross-validation experiments 10 times for each of the three ordering methods (100 trees per method in total). In each experiment, the training set is first randomly divided into ten parts, and the decision tree grown based on nine parts at-a-time, is validated on the remaining one part. The results are given in Table 1 below. The tree sizes and errors on each line are the means of the ten such resulting trees. The errors combine both misclassifications: bents classified as non-bents, and non-bents classified as bents. The astronomers tolerate higher rates of the latter errors, but would like to minimize the mistakes of the former type.

As expected, ordering method III gives the most accurate results. Bent-doubles generally exhibit a symmetry around the core, so this method makes the most sense out of the three considered. We expected method II to

		Decision Tree	
	Exprmt.#	**Size**	**Errors**
Method I.	0	7.4	10.2%
	1	7.0	8.8%
	2	7.5	8.8%
	3	7.4	11.2%
	4	7.0	9.8%
	5	7.7	8.7%
	6	7.2	14.3%
	7	7.1	12.8%
	8	7.1	12.9%
	9	7.2	11.3%
	Mean	7.3	10.9%
	SE	0.1	0.6%
Method II.	0	7.4	9.3%
	1	7.7	9.8%
	2	7.2	12.2%
	3	8.0	10.3%
	4	7.5	10.7%
	5	7.1	12.3%
	6	7.2	14.8%
	7	6.7	14.0%
	8	6.8	11.3%
	9	7.0	13.3%
	Mean	7.3	11.8%
	SE	0.1	0.6%
Method III.	0	6.0	9.7%
	1	6.0	10.8%
	2	5.9	8.7%
	3	6.0	9.8%
	4	5.9	9.7%
	5	6.0	8.2%
	6	5.8	10.2%
	7	6.0	9.7%
	8	5.8	9.8%
	9	5.9	9.8%
	Mean	5.9	9.6%
	SE	0.0	0.2%

Table 1: Results of ten 10-fold cross-validation experiments for the three different ordering methods.

be the next best performer, but, to our surprise, method I gave better results. Our astronomer collaborators indicate that there is no relationship between the flux magnitude and the location of the core, so we are unable to explain this result at present. Selecting the core according to the largest angle, i.e. method II, gave the worst results. We thought it would be superior to method I, as there is greater connection between the geometry of the source and bentness, than there is between the flux and bentness. There are many bent-doubles with the largest angle at the core, so we expected method II to be closer to method III. The latter picks up the two different types of symmetries (core is the largest, or core is the smallest angle), while the former only considers one of the symmetries (core is the largest angle). We are exploring these issues in greater detail in order to fully interpret the results. Note, however, that while the errors are slightly different for the three ordering schemes, they are on the same order. Also, the size of the training set is relatively small at present.

A typical tree constructed with ordering method III is given below.

```
Decision tree:

rs3_core_angle > 170.4:
:...cec_ellipticity <= 2.116: 1 (13.0)
:   cec_ellipticity > 2.116: 5 (2.0)
rs3_core_angle <= 170.4:
:...pairac_rel_dist <= 9.423: 5 (143.0)
    pairac_rel_dist > 9.423:
    :...pairab_angle_geom <= 58.6: 5 (4.0/1.0)
        pairab_angle_geom > 58.6:
        :...cec_rms <= 0.137: 5 (5.0/2.0)
            cec_rms > 0.137: 1 (9.0)

Evaluation on training data (176 cases):

          Decision Tree
        ----------------
        Size      Errors
           6    3( 1.7%)   <<

         (a)   (b)    <-classified as
        ----  ----
          22     3    (a): 1 (non-bent)
               151    (b): 5 (bent)

Evaluation on test data (19 cases):

          Decision Tree
        ----------------
        Size      Errors
           6    2(10.5%)   <<

         (a)   (b)    <-classified as
        ----  ----
           1     2    (a): 1 (non-bent)
                16    (b): 5 (bent)
```

The decision tree output lists the feature selected at each node, as well as the value it is compared against. The number after the colon indicates that the node in question is a leaf node, and the number is the class assigned to the leaf (5 denotes a bent-double, while 1 denotes a non-bent double). At each leaf node, the numbers (a/b) indicate the (total number of samples/samples of the class not assigned to leaf node).

For the 3-entry cases, the decision trees based on this ordering tend to pick combinations of angles and relative distances as the most important features to discriminate bent-doubles. Other features deemed important include measures of ellipticity and symmetry — features that are all scale, rotation, and translation invariant. The angles are usually either the core angle, or pairwise angles calculated geometrically — angles that are robust to small changes in the data. The very reason we included the geometrical angles, **pair_angle_geom**, is exactly to avoid the sensitivity of the differenced angles, **pair_angle_diff**, both explained in Section 3.1.2. The trees generally ignore features related to the fluxes and the areas. Overall, the trees make sense, and they pick the features that we expected in the first place to be closely related to bent-doubleness.

The trees based on the other two ordering schemes were not as consistent as the ones corresponding to the ordering method III. The discriminating features selected occasionally included flux and area measurements, and major axes lengths, in combination with distance and angle values. A few trees that we examined selected actual, rather than relative, distance measurements. The actual distances, and other features such as flux, area, and major axis, are poor discriminating features, as they are not strictly scale invariant. The brightness, and the size of an entry should not be related to bent-doubleness. Our experiments thus indicate that, given the current training set, the ordering methods I and II are inferior to ordering method III in classifying bent-doubles. They have relatively high accuracy, but, on closer examination, they base the classification on features that do not make sense from the domain science point of view, and that keep changing from tree to tree, depending on the training and validation sample selected.

To reduce the number of features, we next repeated the decision-tree building steps, including combinations of the single, double, and triple features. The results for 10 different 10-fold cross-validations for each of the seven combinations, based on the ordering method III are presented in Table 2. The table reinforces our expectation that the most important features are the triple ones. Using only the triple features, the misclassification rate is 10.7%(0.3%), a small increase from the 9.6%(0.2%) achieved when using all the features. The single and/or double features by themselves lead to close to 20% errors. Adding the double features to the triples slightly degrades the results, while adding the singles to the triples slightly improves the results. Characteristic features picked by single+triple combinations include ellipticity, symmetry, relative distance, and angle measurements, while features selected by double+triple combinations consist mainly of distance, area, flux, angle, and ellipticity values. We are not sure what is causing this behavior — it could be a subtle issue, or just chance. We are currently investigating it.

		Decision Tree	
	Value	**Size**	**Errors**
Single	Mean	11.2	19.7%
	SE	0.1	0.5%
Double	Mean	8.7	17.4%
	SE	0.2	0.4%
Single+double	Mean	10.7	19.2%
	SE	0.2	0.5%
Triple	Mean	6.7	10.7%
	SE	0.1	0.3%
Single+triple	Mean	6.4	8.5%
	SE	0.0	0.4%
Double+triple	Mean	7.1	11.6%
	SE	0.1	0.5%
Single+double+triple	Mean	5.9	9.6%
	SE	0.0	0.2%

Table 2: Average of ten 10-fold cross-validation experiments for each of the seven 3-entry feature combinations.

An early version of our decision tree, when used for classifying unlabeled data, found several new bent-doubles, as expected. For example, Figure 6 shows an example of a new bent-double from the region the astronomers had not looked at manually (left panel). What is interesting is that the data mining process also found a bent-double that the astronomers had missed (right panel) during the visual inspection that generated the training set. This illustrates some of the many benefits of data mining techniques in the semi-automated exploration of massive data sets.

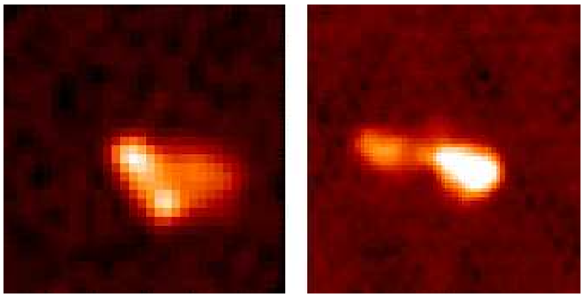


Figure 6: Examples of new bent-doubles.

5 Summary

In this paper, we described how data mining techniques can help astronomers detect radio galaxies with a bent-double morphology in a semi-automated manner. Our experiences indicate that the identification and extraction of relevant features plays a very important role in the accuracy of the pattern recognition algorithms. Though much remains to be done, our initial results appear very promising. Our immediate plans for the bent-double problem include increasing the size of the training set, revising the catalog-based features, and adding image-based features. Revising the catalog-based features has been an ongoing process. For the triple sources, our average misclassification rate of about 10% is half the rate we initially obtained during the first iteration of the data mining process. New features, such as the angles ACN and ABM in Figure 5, keep emerging as we discuss our findings with our astronomer collaborators. They expect that the smaller these angles, the more likely a triple source is a bent-double. Another potential way to improve the features derived from the catalog is to remove possible redundancies among the various angle and distance measurements by combining them into fewer, more relevant features. In addition, as we write our own decision tree software, we are interested in seeing how the accuracy of the trees generated for the bent-double problem will change as we change the split criteria, or use oblique decision trees induced by hill climbing, randomization, or evolutionary algorithms (Cantú-Paz and Kamath, 2000). We also plan on using other pattern recognition techniques such as neural networks to see how they perform on the bent-double problem.

Acknowledgments

We gratefully acknowledge our FIRST collaborators Robert H. Becker, Michael D. Gregg, David J. Helfand, Sally A. Laurent-Muehleisen, and Richard L. White for their technical interest and support of this work. We would also like to thank Charles R. Musick, Deanne D. Proctor, and Ari Buchalter for useful discussions and/or computational help.

UCRL-JC-138073. 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.

References

Becker, R.H., R.L. White, and D.J. Helfand (1995) "The FIRST Survey: Faint Images of the Radio Sky at Twenty-cm," *Astrophysical Journal*, Vol. 450, p. 559.

Cantú-Paz, E. and C. Kamath (2000) "Using Evolutionary Algorithms to Induce Oblique Decision Trees", *Genetic and Evolutionary Computation Conference (GECCO) 2000*, Las Vegas, NV, July 2000.

Kamath, C., and R. Musick (2000) "Scalable Data Mining through Fine-Grained Parallelism: The Present and the Future," accepted for publication in *Advances in Distributed Data Mining*, H. Kargupta and P. Chan, Eds., to be published by AAAI Press, Summer 2000.

Lehar, J. et al. (2000), *In Preparation.*

White, R.L., R.H. Becker, D.J. Helfand, and M.D. Gregg (1997) "A Catalog of 1.4 GHz Radio Sources from the FIRST Survey," *Astrophysical Journal*, Vol. 475, p. 479.

White, R.L. (1999) "Private Communication".