Chapter 1

SEARCHING FOR BENT-DOUBLE GALAXIES IN THE FIRST SURVEY [†]

Chandrika Kamath, Erick Cantú-Paz, Imola K. Fodor, and Nu Ai Tang

Abstract

Data mining techniques are increasingly gaining popularity in various scientific domains as viable approaches to the analysis of massive data sets. In this chapter, we describe our experiences in applying data mining to a problem in astronomy, namely, the identification of radio-emitting galaxies with a bent-double morphology. Until recently, astronomers associated with the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey identified these galaxies through a visual inspection of images. While this manual approach has been very subjective and tedious, it is also becoming increasingly infeasible as the survey has grown in size. Upon completion, FIRST will include almost a million galaxies, making the use of semi-automated analysis methods necessary. We describe the FIRST data set and the problem of identifying bent-double galaxies. We discuss our solution approach, focusing on the challenges we face in the application of data mining to a scientific data set. We explain why, in contrast with most commercial data mining applications, data preprocessing requires a considerable effort in scientific applications. Using decision tree classifiers, we describe the work we are doing in the detection of bent-double galaxies. Our results indicate that data mining techniques, steered by proper domain knowledge, can greatly enhance the manual exploration of massive data sets.

Keywords: Data mining, astronomical surveys, radio-emitting galaxies, role of data preprocessing

1. Introduction

Data mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures and events in data ([KM00]

[*] To be published in Data Mining for Scientific and Engineering Applications, Kluwer 2001

and the references therein). It is an iterative and interactive process involving data preprocessing, search for patterns, and interpretation of the results. While there is broad consensus on what constitutes data mining, the tasks that are performed in each step depend on the problem domain, the problem being solved, and the data itself. As explained later, scientific datasets present particular challenges that require a careful tailoring of the data mining tasks to the problem and dataset.

In this chapter, we describe how the Sapphire project [Sap] has defined the tasks in data mining to make the process more suitable for addressing the diverse needs of scientific and engineering applications. We discuss our work in the context of one of the data sets we are analyzing from astronomy, where we are interested in identifying radio-emitting galaxies with a bent-double morphology. Our goal in this effort is to replace the current visual inspection of images by a semi-automated, and hopefully more objective, approach based on data mining techniques. We discuss the challenges we have faced in trying to reach this goal, and the approach we have taken to address some of these problems. In particular, we show that data preprocessing is not only crucial to the successful application of data mining, but can be a difficult and time consuming task for scientific data.

This Chapter is organized as follows: Section 2 describes our view of scientific data mining, focusing in particular on the tasks that are specific to scientific and engineering data sets. Section 3 describes our astronomy data set, namely the FIRST survey, and outlines the problem of detecting bent-double radio-emitting galaxies. Section 4 provides details on the approach we have taken to mine the FIRST data set for bent-doubles, and the difficulties we have faced in the process. Section 5 reports our results, focusing on the important role played by the data preprocessing step in data mining. Section 6 concludes with our observations, the lessons we have learned, and plans for future work.

2. Scientific Data Mining

Data sets that arise in various physical sciences and engineering applications are usually the result of computer simulations, observations, or experiments. These data are typically in the form of signals (including images) or mesh data. Often, one or more types of data may be available for a problem, in which case data fusion is required in order to exploit all the information and make a more accurate decision.

In light of these observations, our definition of data mining starts with the raw data and includes extensive preprocessing (Figure 1.1). If the raw data is very large, we may use sampling and work with fewer instances, or use multiresolution techniques and work with data at a coarser resolution. This first step may also include data fusion, if required. Next, noise is removed and

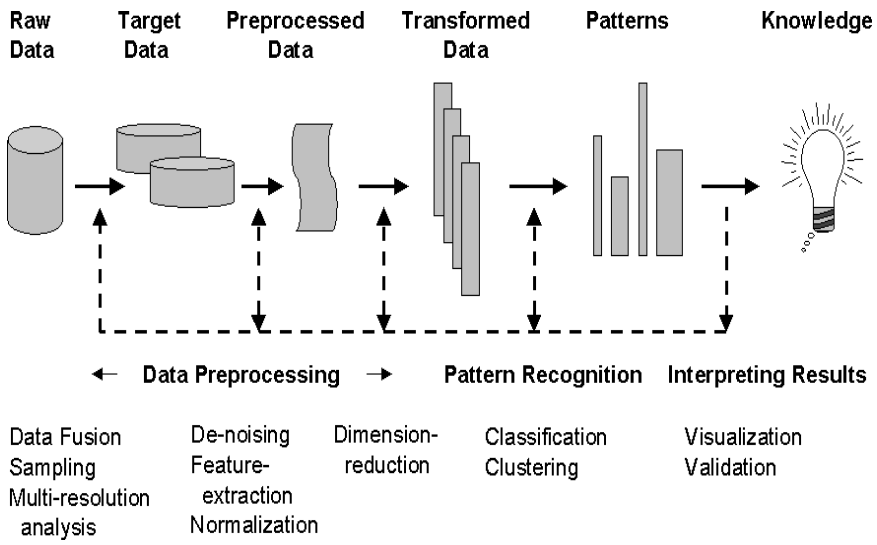


Figure 1.1. Data mining: an iterative and interactive process.

relevant features that represent and describe the items of interest are extracted from the data. These features must not only be robust, that is, insensitive to small changes in the data, but also be invariant to scaling, translation, and rotation. At the end of this step, we have a vector of features for each data item. These vectors are normalized to remove any skew resulting from possibly different units. Next, depending on the problem and the data, we may need to reduce the number of features using dimension reduction techniques such as principal component analysis (PCA) or its non-linear versions. After this preprocessing, the data is ready for the detection of patterns, through the use of traditional techniques such as classification and clustering. The patterns obtained are then displayed to the domain scientist for validation.

The data mining process is iterative and interactive; any step may lead to a refinement of one or more of the previous steps. To ensure the success of data mining, the domain scientist should be actively involved in all stages, starting from the initial description of the data and the problem, the identification of potentially relevant features and the training set (where necessary), and the validation of the results.

It must be noted that several of the tasks in scientific data mining are independent of the data or the problem, while others are more specific to the domain. However, for scientific data in the form of signals, images, or meshes, there is ample opportunity for re-use of algorithms and software. For example, while wavelet-based de-noising techniques can be used in general to de-noise signals and images, the particular type of wavelet used might depend on the statistical

properties of the data and noise, which in turn is determined by the sensors that were used to collect the data. Similarly, while decision trees may be used in the classification of both astronomy and simulation data, it may turn out that neural networks are better at classification for a particular problem in astronomy. In addition, techniques that may have been developed for one kind of data, such as segmentation for images, can be applied to a similar task in another kind of data, such as segmentation in mesh data. However, tasks such as reading, writing, and displaying data are typically very specific to the format used for the data.

As part of the Sapphire project, we are developing toolkits for each of the computationally intensive modules described above. This object-oriented framework, written in C++, is designed to be modular, with the ability to handle both signal and mesh data types. Further details on this project can be found in [Sap, KCP00, KBFT00].

We next describe how we are applying the data mining process described in this Section to a problem in astronomy.

3. The FIRST Survey

The FIRST (Faint Images of the Radio Sky at Twenty-cm) survey [BWH95] 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 1999 observations, FIRST has covered about 8,000 square degrees, producing more than 32,000 two-million pixel images. At a threshold of 1mJy, there are approximately 90 radio-emitting galaxies, or radio sources, in a typical square degree. The results we present in this chapter are based on the 1998 version of the catalog, which includes data from 1993 through 1997.

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. FIRST scientists currently use a manual approach to detect bent-double galaxies. 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.

This visual inspection of the radio images, besides being very subjective, is also becoming increasingly infeasible as the survey grows in size. Our goal is to

bring automation to this process of classifying galaxies by means of techniques from data mining.

Figure 1.2 includes several examples of radio sources from the FIRST survey. Radio galaxies can have rather complex shapes, as shown in examples (g) through (l). While some bent-doubles are relatively simple in shape (examples (b) and (c)), others, such as the ones in Figure 1.3 can be rather complex. The task of automating the detection of bent-doubles can be quite complex as seen from the similarity between the bent-double in example (b) and the non-bent-double in example (d).

The data from FIRST, both raw and postprocessed, are readily available on the FIRST website [FIR]. 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. In Figure 1.3, 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 an area approximately 0.45 square degrees, with pixels that are 1.8 arc seconds wide. These image maps are obtained as a result of processing the raw data collected by the VLA telescopes.

In addition to the image maps, the FIRST survey also provides a source catalog [WBHG97]. This catalog is obtained by processing an image map by fitting two-dimensional elliptic Gaussians to each radio source. For example, the lower bent-double in Figure 1.3 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). Each of the three entries in the catalog corresponds to one of the three "blobs" in the image. Note that we differentiate between catalog entries and radio sources, with a radio source being composed of one or more catalog entries.

4. Identification of Bent-Doubles

As illustrated in Figure 1.3, we have data at two extremes. On one hand, we have 200 Gigabytes of image maps, which are mainly noise, with very few "interesting" pixels corresponding to the radio sources. On the other hand, we have the 78 Megabyte catalog information, where each entry contains information on only a part of a radio source. The first task therefore is to identify what constitutes a radio source, using either the image maps, or the catalog, or

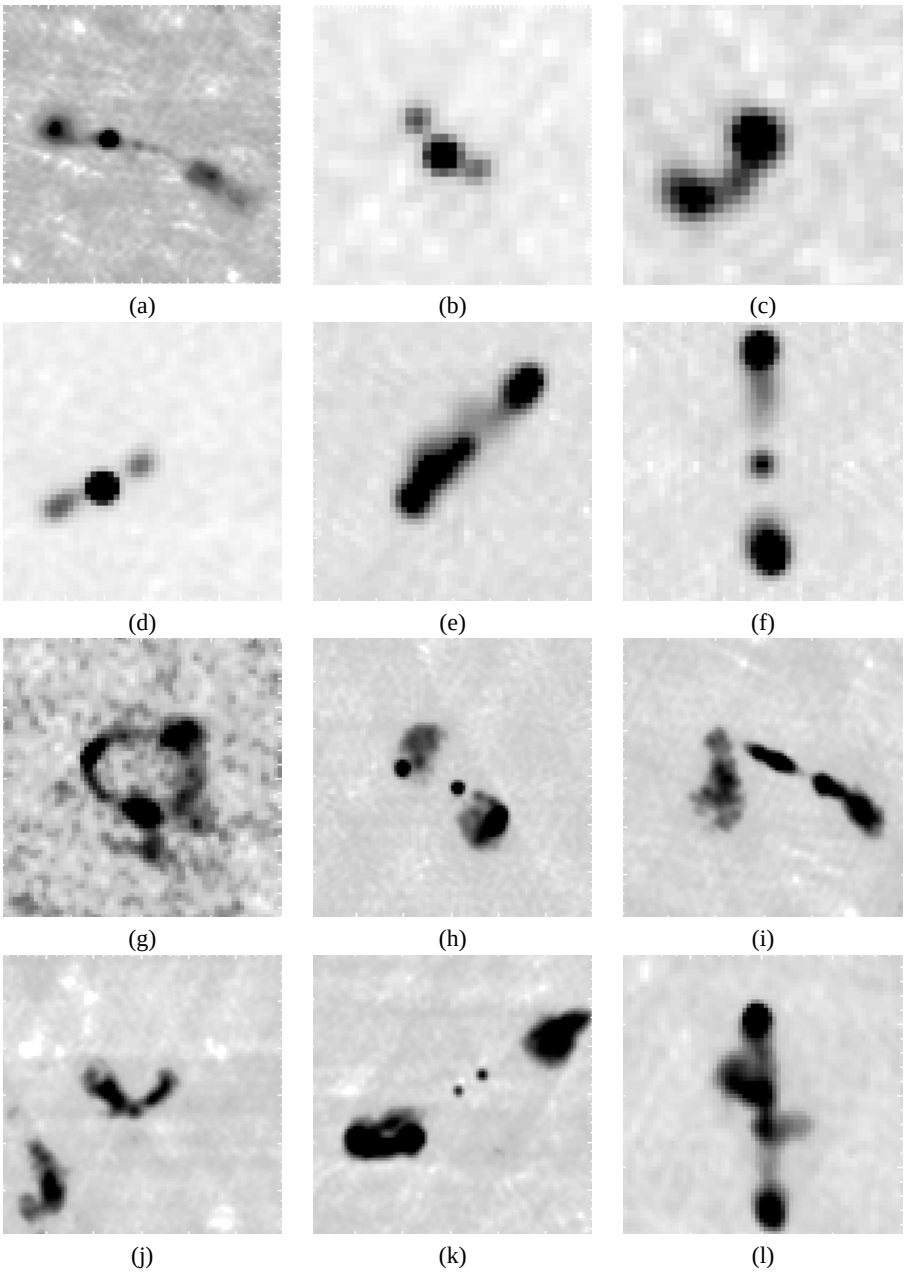
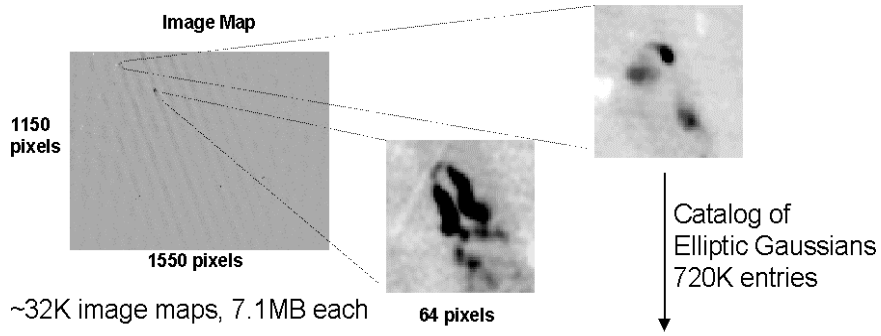


Figure 1.2. Example radio sources from FIRST (a)-(c) Bent-doubles. (d)-(f) Non-bent doubles (g)-(l) Complex Sources



Catalog entry

Radio Galaxy

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 1.3. FIRST Data: Images Maps and Catalog Entries.

both. Once we have done this, we need to first identify the potentially relevant features for each galaxy, extract them, and then use them in the identification of the bent-doubles.

In our work, we decided that initially, we would identify the radio sources and extract the features using only the catalog. This choice was prompted by several factors:

- The astronomers believed that the catalog was a good approximation to all but the most complex of radio sources.
- It was easier for us to work with the catalog as it was smaller.
- Processing the relatively 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 Section 5, we comment on the effects of our decision to use only the information in the catalog.

Since we decided to work with the catalog, our first task 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 new 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 any 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 the center of the new entry, or the center of mass of the entries that made 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 1.4 with the image centered at RA = $10^{\mathrm{h}}50^{\mathrm{m}}08.5^{\mathrm{s}}$ and Dec = $+30^{\mathrm{o}}40'15''$ (J2000 coordinates), shows two radio sources, a bent-double in the lower left corner, and a ring-like structure, which is also a bent-double in the upper right corner. While these radio sources may be far from each other in three dimensions, in a two-dimensional projection, they appear close together. Such examples illustrate one of the many reasons why the task of automated detection of bent-doubles is a rather hard problem. It also shows the ease with which humans can visually identify the two objects as being separate, a task that is difficult to automate.

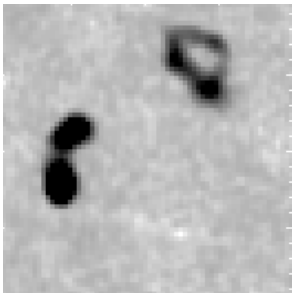


Figure 1.4. An example image from FIRST, illustrating two galaxies close together

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. There is a data set each for the 1-entry sources, the 2-entry sources, the 3-entry sources, and the 3-plus-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 3-plus-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. However, it did not address the case where two disconnected sources, close to each other, were approximated by two or three Gaussians.

Having removed the 1-entry and the 3-plus-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, 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 4.1 for the two and three entry sources. Next, using the appropriate training set, we created decision trees for the identification of two- and three- entry radio sources. We ran cross-validation experiments to estimate the accuracy of the tree using different subsets of the features.

4.1. Features for Bent-Doubles

This section describes the features we are using to discriminate galaxies with bent-double morphology. Some of the features are directly taken from the FIRST catalog and others are derived from the basic features in the catalog. We also include a few "features", such as the radio source ID and position in the sky, for bookkeeping purposes only.

Our focus is on features that are scale, rotation and translation invariant, as the bent-double pattern we are looking for, has these properties. We are also interested in features that are robust, that is, not sensitive to small changes in the data [Whi99]. 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 some of the key features we calculated based on our collaboration with the FIRST astronomers. A full list of features is described in [FCPKT00]. We broadly categorize the features based on the number of catalog entries that are used in their calculation.

- Features for the radio source

  This includes features that pertain to the entire radio source, such as the number of catalog entries, or bookkeeping features such as the radio source ID (for tracking purposes), or the hemisphere for the radio source (nothern or southern).

- Features for each catalog entry

  This includes features pertaining to a single catalog entry, such as its peak flux, total area of the elliptic Gaussian, the ellipticity of the Gaussian, the major and minor axes, etc. We also include the position angle, which is the angle (in degrees) of the major axis, measured counterclockwise from North. In Figure 1.5, the angles are indicated by an arrow — for entry B it is about 45^o in the left example, and about $(180 - 45)^o$ in the right example. The angle is 0 for entry A in both examples. Note that the position angle is not a robust feature as it is very sensitive to minor changes in the image.

- Features for each pair of catalog entries

  This category includes two types of features. In the case of two-catalog-entry radio sources, these are features representing the radio source, such as the total area (the sum of the areas of the two Gaussians) and the peak flux (the larger of the two fluxes). In addition, there are features that are obtained by considering catalog entries two at a time. These include relative distance between two entries, and the pair-wise geometric angle, which is the angle formed by the position angles of the two major axes, as calculated geometrically – angle AMB in both examples of Figure 1.5. We also include the absolute difference in the position angles of the two entries, which is about $|0 - 45|^o = 45^o$ in the left example, and about $|0 - 135|^o = 135^o$ in the right example of Figure 1.5. Note that this feature is not robust to small changes in the image.

- Features for each triple of catalog entries

  In the case of three-entry sources, some features, such as the total area (sum of the areas of the Gaussians), represent the entire radio source.

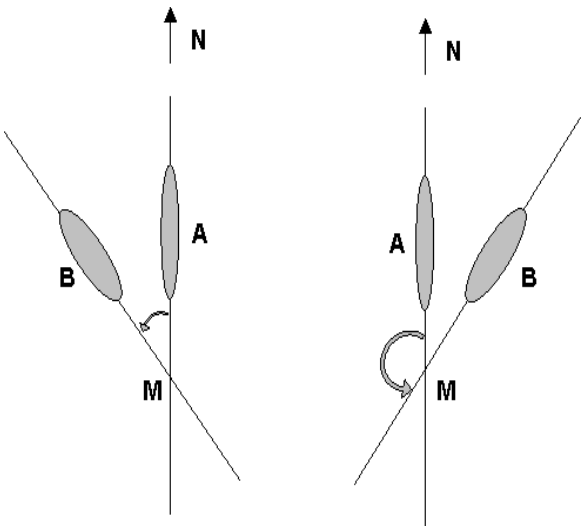


Figure 1.5. Two examples of 2-entry fitted radio sources.

Others include various measurements of angles, such as the angle subtended by the largest side of the triangle created by the centers of the three Gaussians [LBR$^+$99], and the angle subtended by the two sides of the triangle that are closest to each other in legnth.

In the case of two-entry sources, the features were listed in decreasing order of integrated flux for each Gaussian. The features include the ones for the radio source, the ones for each catalog entry and the ones for the pair of catalog entries. In the case of 3-entry radio sources, we include all four categories of features described earlier. We also explored three ways of ordering the three entries. These were based on first identifying one of the catalog entries as the "core" of the radio source. The options we considered include:

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 some-what 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.

5. Results Using Decision Trees

In this Section, we present some of our early results using decision tree classifiers [BFOS84, Qui93] and the features from the catalog described in Section 4.1. We first make the following observations:

- Our training set is relatively small, with 118 examples for two-catalog entry sources and 195 examples 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. We plan to enhance our small training set by using feedback from the astronomers on the results of the preliminary decision trees.

- Scientists are usually subjective in their labeling of galaxies as bent- or non-bent-doubles. There is often disagreement among the astronomers in the hard-to-classify cases. There is also no ground truth we can use to verify our results. This implies that the training set itself is not very accurate, and there is a limit to the accuracy we could obtain through the use of semi-automated techniques.

- 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.

In our experimentation, 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

Method	*Tree Size*	*Errors*
Method I.	7.3 (0.1)	10.9 (0.6)%
Method II.	7.3 (0.1)	11.8 (0.6)%
Method III.	5.9 (0.0)	9.6 (0.2)%

Table 1.1. Results of ten 10-fold cross-validation experiments for the three different ordering methods.

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.

5.1. Results for 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 4.1, 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 1.1. The tree sizes and errors on each line are the means of the hundred trees, with the standard error in parenthesis. 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 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, these results must be interpreted keeping in mind that 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 ordering III 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 these pairwise geometrical angles (Section 4.1), was to avoid using the more sensitive feature that measured the absolute difference in the position angles. 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 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 initial 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, using 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 1.2.

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 improves the results. This behavior may be due to the existence of redundant features; it is currently under investigation.

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

Table 1.2. 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 1.6 shows two examples of new bent-doubles from the region the astronomers had not looked at manually (left panels). 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 one of the many benefits of data mining techniques in the semi-automated exploration of massive data sets.

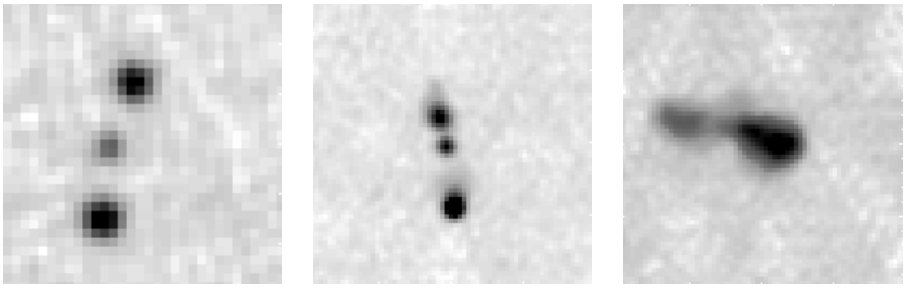


Figure 1.6. Examples of two new bent-doubles and one bent-double overlooked in a manual search

5.2. Results for Two-Entry Sources

Our initial experiments with two-entry sources found that the observations made in Section 5 play an important role. 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%.

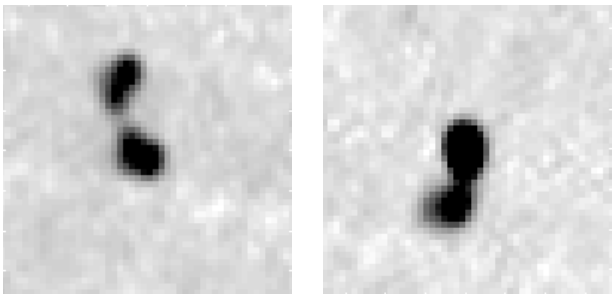


Figure 1.7. Examples of 2-entry galaxies - a bent-double and a non-bent-double.

We suspect there are several reasons for this, including the small training set (118 examples, with 72 bent-doubles and 46 non-bent-doubles) and possible irrelevant features. In addition, we suspect that in the case of two-entry sources, the catalog may not be an adequate representation of all the relevant information required to correctly identify a bent-double. For example, Figure 1.7 includes two entries from our training set, one a bent-double and the other a non-bent-double. Some astronomers have indicated that the faint bridge in the left example (which is not captured in the catalog) makes it a bent-double. In addition, these examples fall in the hard-to-classify case, making our training set not very accurate.

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.

6. Conclusions and Future work

In this chapter, we described how data mining techniques can help astronomers detect radio galaxies with a bent-double morphology in a semi-automated manner. While we can by no means claim to understand and appreciate all the challenges that lie in mining scientific and engineering datasets, our experiences with this problem has led to the following observations:

- Scientific data is frequently available in the form of "raw" data, that is, data from which features have not been extracted. In our experience, the identification and extraction of relevant features in a robust manner is non-trivial and results in one of the more time consuming steps in data mining. In our specific example of the detection of bent-doubles, the existence of the catalog was of immense help in getting a head-start on the problem.

- Easy access to the data, as well as the existence of good public-domain software to read, write, and display the data are essential in order for data miners to work with scientific data sets. In the case of the FIRST data, the availability of the data on the web, as well as tools to read/write the astronomy data stored in the FITS format, were very helpful.

- In many, if not most, data mining endeavors, it takes some time for the data miners to understand the problem domain, the problem, and the data itself. As many data miners are typically not familiar with the intricacies of the issues involved in scientific data, as opposed to say commercial data, this time can be considerable. In our experience, we found it took six months to understand the bent-double problem, the data formats, the processing that had already been done to the data, etc.

- In scientific data, for classification problems, there can be a dearth of labeled examples. This is because such examples must be manually identified by the scientists. This is in contrast with commercial data, where there is the possibility of labeled examples being generated historically, for example, in customer churn problems. Further, labels for scientific data are typically subjective, and we found it common to have astronomers disagree on whether a galaxy should be classified as a bent-double or a non-bent-double. This was especially true in the cases which were difficult to classify. Given the lack of ground truth in this problem, generating a good training set has been difficult.

- Since FIRST is a survey in progress, we found that our bookkeeping had to keep up with changes made in different releases of the data. For example, the 1999 version of the data merged the information from the northern and southern hemisphere, information that had previously been 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 that was collected by the telescopes. For galaxies that were at the very edge of the survey in one year, additional data collected in the following year would make the pixels corresponding to the 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 move to newer versions of the survey.

Our initial experiences with the bent double problem appear promising, though much remains to be done. In the near term, we plan on 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 three-entry 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 emerge as we discuss our findings with our astronomer collaborators. We are also improving the features derived from the catalog by removing possible redundancies among the various angle and distance measurements by combining them into fewer, more relevant features. 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 Becker, Michael Gregg, David Helfand, Sally Laurent-Muehleisen, and Richard White for their technical interest and support of this work. We would also like to thank Charles Musick, Deanne Proctor, and Ari Buchalter for useful discussions and/or computational help.

UCRL-JC-140418. 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

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* CRC Press, 1984.

[BWH95] R. H. Becker, R.L. White, and D.J. Helfand. The FIRST survey: Faint images of the radio sky at twenty-cm. *Astrophysical Journal,* 450:559, 1995.

[FCPKT00] 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, April 2000.

[FIR] FIRST: Faint images of the radio sky at twenty centimeters. http://sundog.stsci.edu/.

[KBFT00] C. Kamath, C. Baldwin, I. Fodor, and N. Tang. On the design and implementation of a parallel, object-oriented, image processing toolkit. In *Proceedings International Symposium on Optical Science and Technology, SPIE Annual Meeting, San Diego,* July 2000.

[KCP00] C. Kamath and E. Cantú-Paz. On the design of a parallel object-oriented data mining toolkit. In *Workshop on Distributed and Parallel Knowledge Discovery at the Knowledge Discovery and Data Mining Conference Boston,* August 2000.

[KM00] C. Kamath and R. Musick. Scalable data mining through fine-grained parallelism: The present and the future. In H. Kargupta and

P. Chan, editors, *Advances in Distributed and Parallel Knowledge Discovery*, pages 29–77. AAAI Press/The MIT Press, 2000.

[LBR+99] J. Lehar, A. Buchalter, R.McMahon, C.Kochanek, D.Helfand, R.Becker, and T.Muxlow. The FIRST efficient gravitational lens survey. 1999. submitted to "Gravitational Lensing: Recent progress and Future Goals, eds: T.Brainerd and C.Kochanek, ASP Conf Series See also http://xxx.lanl.gov/abs/astro-ph/9908353.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufman, 1993.

[Sap] Sapphire: Large-scale data mining and pattern recognition. http://www.llnl.gov/casc/sapphire.

[WBHG97] R. L. White, R.H. Becker, D.J. Helfand, and M.D. Gregg. A catalog of 1.4 GHz radio sources from the FIRST survey. *Astrophysical Journal*, 475:479, 1997.

[Whi99] R. L. White, 1999. Private Communication.