Preprint
UCRL-JC143458

Using Data Mining to Find Bent-double Galaxies in the FIRST Survey

C. Kamath, E. Cantu-Paz, I. K. Fodor, N. Tang

This article was submitted to Astronomical Data Analysis at the SPIE's International Symposium on Optical Science and Technology, San Diego, CA, July 29-August 3, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

June 22, 2001

Using Data Mining to Find Bent-Double Radio Galaxies in the FIRST Survey

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

Center for Applied Scientific Computing, Lawrence Livermore National Laboratory
P.O. Box 808, L-561, Livermore, CA 94551

ABSTRACT

In this paper, we describe the use of data mining techniques to search for radio-emitting galaxies with a bent-double morphology. In the past, astronomers from the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey identified these galaxies through visual inspection. This was not only subjective but also tedious as the on-going survey now covers 8000 square degrees, with each square degree containing about 90 galaxies. In this paper, we describe how data mining can be used to automate the identification of these galaxies. We discuss the challenges faced in defining meaningful features that represent the shape of a galaxy and our experiences with ensembles of decision trees for the classification of bent-double galaxies.

Keywords: Data mining, astronomical surveys, radio-emitting galaxies

1. INTRODUCTION

Data mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures and events in data ([1] and the references therein). It is an iterative and interactive process involving data preprocessing, search for patterns, and interpretation of the results (see Figure 1). Input from domain scientists is an integral part of the data mining process, and frequently results in the refinement of one or more steps. 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.

This paper describes our results in applying data mining techniques to data from the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey. We are interested in identifying radio-emitting galaxies with a bent-double morphology. Previously, we obtained promising results using decision trees to classify galaxies according to their morphology.[2] Decision trees are popular classification algorithms since they are easy to understand, can be built efficiently, and usually have good accuracy.[3,4] To attempt to improve on the accuracy of single decision trees, we can combine several decision trees into ensembles, which combine the output of several classification algorithms and are known to be more accurate than individual trees. This paper presents recent results using three types of ensembles of decision trees.

The paper is organized as follows: Section 2 describes the data set from the FIRST survey, and outlines the problem of detecting bent-double radio-emitting galaxies. Section 3 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 4 reports our results. Section 5 concludes with our observations and plans for future work.

2. THE FIRST SURVEY

The FIRST (Faint Images of the Radio Sky at Twenty-cm) survey[5] 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.

Further author information: (Send correspondence to C. Kamath)
C.K.: E-mail: kamath2@llnl.gov

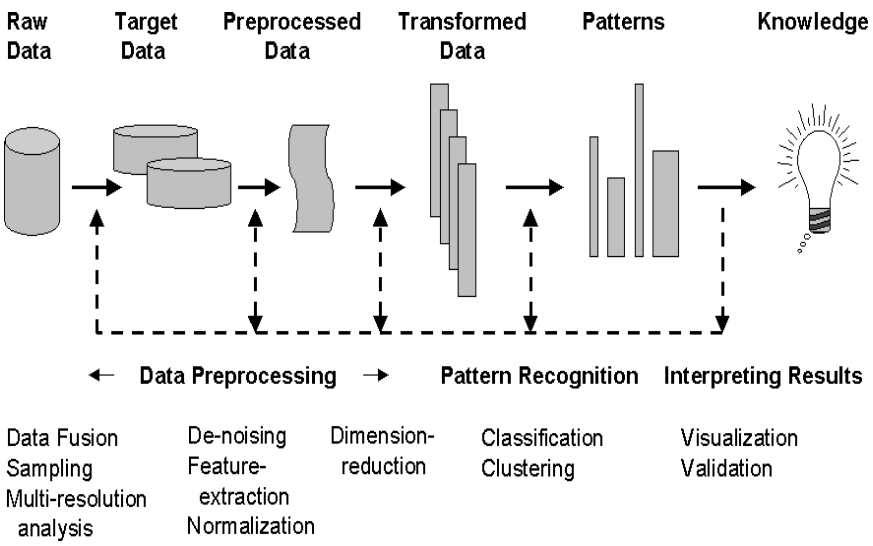

Figure 1. Data mining: an iterative and interactive process.

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 2 includes several examples of radio sources from the FIRST survey. While some radio sources are relatively simple in shape (examples (b) and (c)), others, such as the ones in examples (g) and (h), can be rather complex. The task of automating the detection of bent-doubles can be quite difficult 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.[6] 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 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.[7] 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 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.

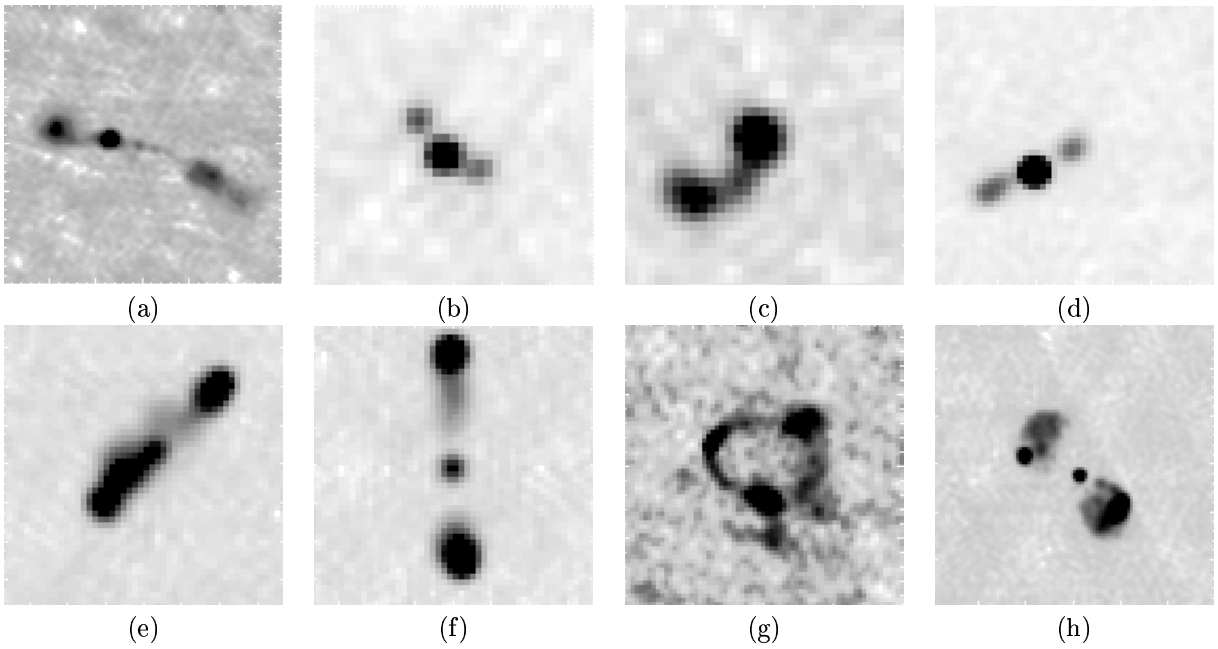

Figure 2. Example radio sources from FIRST: (a)-(c) Bent-doubles, (d)-(f) Non-bent doubles, (g)-(h) Complex Sources

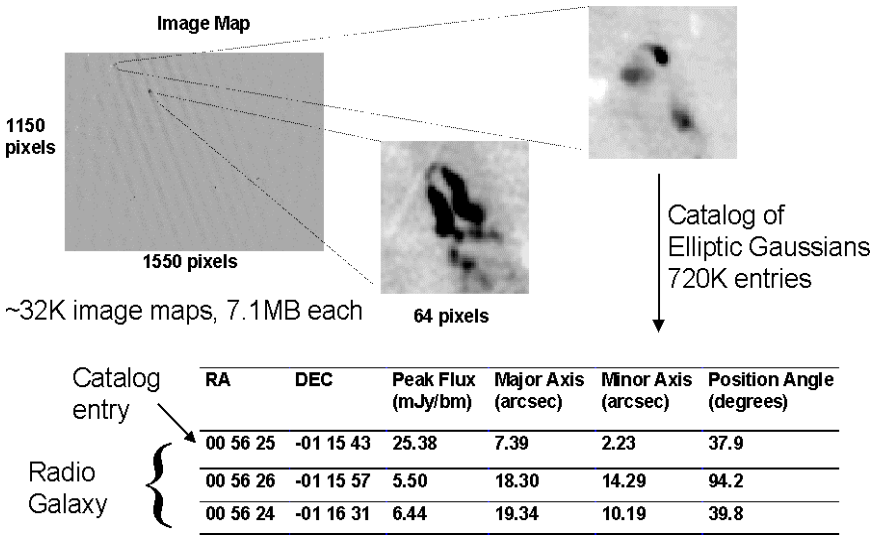

Figure 3. FIRST Data: Images Maps and Catalog Entries.

3. IDENTIFICATION OF BENT-DOUBLES

As illustrated in Figure 3, we have data at two extremes. On one hand, we have 250 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 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 much smaller than the image maps.
- Processing the relatively large image maps for extracting relevant features for the bent-double problem was expected to be difficult and time consuming given the variability in the images.
- The FIRST astronomers indicated that several of the features they thought were important in identifying bent-doubles were easily calculated from the catalog.

Since we decided to work with the catalog, our first task in classifying the bent-doubles 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 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 4 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, 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.

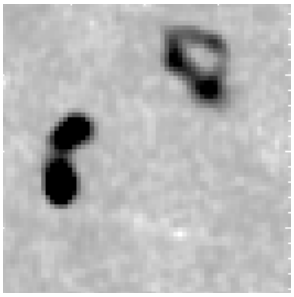

Figure 4. An example image from FIRST, illustrating two galaxies close together

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 2000 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	514637
2	66571
3	15059
3+	6333

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 the next section for the two and three entry sources.

3.1. 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.[8] 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 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 elsewhere.[9] 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 (northern 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 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

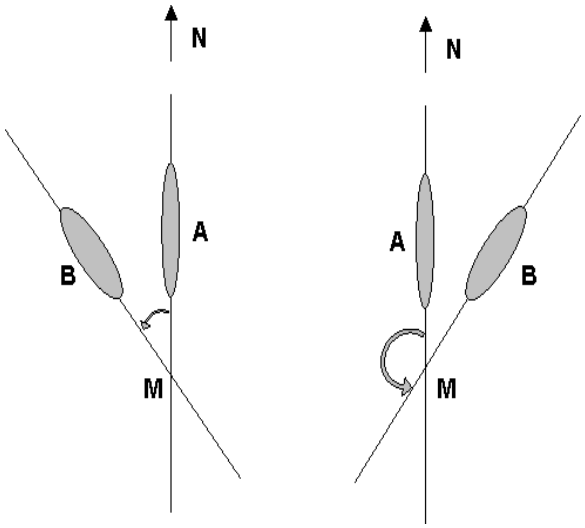

Figure 5. Two examples of 2-entry fitted radio sources.

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 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 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. 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,[10] and the angle subtended by the two sides of the triangle that are closest to each other in length.

In the case of 3-entry radio sources, we include all four categories of features described above. We also explored three ways of ordering the three entries. These orderings were based on first identifying one of the catalog entries as the "core" of the radio source. Our previous experiments with this data suggest that the best ordering is to choose the core to be the entry opposite the side that is most "unlike" the other two sides. The entries are in the following order: 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). Bent-doubles generally exhibit a symmetry around the core, so this method makes the most sense out of the three considered.

Once the features are extracted, they can be used as inputs to a classifier and refined until the accuracy required by the astronomers is reached. We can use any classifier (decision trees, neural networks, ensembles, etc.), but we expect that the results are much more dependent on the features than on the particular classifier that we use. We have used decision trees before with promising results.[2] The next section presents recent results with ensembles of decision trees.

4. ENSEMBLES OF DECISION TREES

An ensemble of classifiers is a set of classifiers whose outputs are combined in some way to classify unseen instances. Typically, ensembles are combined by a (possibly weighted) majority voting scheme. There is ample empirical evidence[11–13] that ensembles are more accurate than individual classifiers. The main drawback of ensemble learning is that the computational cost increases proportionally to the number of classifiers that form the ensemble.

In this section we present the results of experiments with three ensemble methods: bagging, AdaBoost, and ArcX4. Bagging was introduced by Breiman[14] and consists of presenting the tree-building algorithm with a training set that consists of a sample of n examples drawn randomly with replacement from the original training set of n examples. On average, each sample contains 63.2% of the original training examples, with several examples appearing multiple times. The training set is resampled every time before building a tree. The second ensemble method that we considered was AdaBoost, which was developed by Freund and Schapire.[15] This method maintains a weight associated with each example in the training set. The initial weights are set uniformly to 1. At each iteration, i, the algorithm builds a tree using the entire training set and adjusts (boosts) the weights of examples that are misclassified by the tree by multiplying by $\beta_i = \epsilon_i/(1-\epsilon_i)$, where ϵ_i is the sum of the weights of the misclassified examples in the i-th iteration, divided by the size n of the training set. After boosting, the weights are normalized so they add to n. The votes in the final classifier are weighted according to the accuracy of each tree on its weighted training set. The third ensemble method that we tried was ArcX4 that was developed by Breiman[16] and, similarly to AdaBoost, creates classifiers based on weighted training sets. In ArcX4 the weights are set to $1 + e(x)^4$, where $e(x)$ is the number of misclassifications made on instance x by all the previous classifiers. The weights are normalized so that the total weight is equal to the size of the training set. The vote of the final classifier is not weighted in ArcX4.

Before presenting our results using ensembles of decision trees and the features from the catalog described in Section 3.1, we make the following observations:

- Our training set is relatively small, with 195 examples for the three-catalog entry sources (167 bent-doubles and 28 non-bent-doubles). 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.

Using all the features listed in Section 3.1, we estimated the generalization error using standard 10-fold cross-validation experiments. In each experiment, the training set is first randomly divided into ten parts, and the ensemble produced based on nine parts at-a-time, is validated on the remaining one part. In the experiments we varied the number of trees used in each ensemble. The results are given in Table 1. We report the average error in classification and the standard error for each experiment. 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, the experiments show that the error generally decreases as more trees are used in the ensembles. Bagging consistently gives the most inaccurate ensembles, although the differences are not significant. The crossvalidated estimate of the error of a single unpruned tree trained on the entire data was 0.1263 (with 0.2131 standard error). The ensembles with 10 trees are at least as good as the single trees, and ensembles with 20 trees or more are consistently better than single trees. For our data, an ensemble of 20 trees produced by AdaBoost seems a good choice, because using more trees with AdaBoost does not improve the accuracy but increases the cost of making the ensemble. A slightly better result was obtained by an ensemble of 100 trees produced by ArcX4, but the difference is not significant and the cost is much higher.

Our previous experience with this data suggests that we can find more accurate decision trees if we limit the training set to contain only the features for triples of catalog entries.[9] Table 2 contains the results of experiments

	Bagging		AdaBoost		ArcX4	
Trees	Error	Std error	Error	Std error	Error	Std error
2	0.1315	0.0213	0.1315	0.02	0.121	0.0223
5	0.1368	0.0237	0.1105	0.0157	0.1315	0.0213
10	0.1263	0.0225	0.1052	0.021	0.1052	0.0182
20	0.1052	0.0235	0.0894	0.0247	0.1	0.0173
50	0.1052	0.0196	0.0894	0.0247	0.1052	0.0182
100	0.1105	0.0203	0.0894	0.0247	0.0842	0.0199

Table 1. Ten-fold crossvalidation results using different ensemble methods on all the features.

	Bagging		AdaBoost		ArcX4	
Trees	Error	Std error	Error	Std error	Error	Std error
2	0.1263	0.0237	0.121	0.0316	0.1	0.0203
5	0.1157	0.0314	0.0684	0.0106	0.0789	0.02
10	0.121	0.0197	0.0789	0.017	0.0684	0.0149
20	0.0894	0.0167	0.0842	0.0199	0.0789	0.0186
50	0.0947	0.0179	0.0526	0.0166	0.0894	0.0211
100	0.0947	0.0194	0.0736	0.0213	0.0947	0.0207

Table 2. Ten-fold crossvalidation results using different ensemble methods on the triple features.

with ensembles on the triple features. In general, the error rate is lower in this case than with all the features, but the differences are not significant. As in the previous experiments, we observe that very small ensembles have higher error rates than the larger ensembles, and bagging had slightly less accurate results than the other methods. As before, it seems wasteful to use large ensembles on this data, since we obtained good results with ensembles with 5–10 trees. In fact, using larger ensembles resulted in a higher error rate, which may be attributed to overfitting.

5. CONCLUSIONS AND FUTURE WORK

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. After extracting the features that describe the galaxies, we performed experiments with different ensembles of decision trees. Our results are in agreement with previous experimental studies of ensembles that indicate that bagging usually produces more inaccurate ensembles than boosting methods. We found that relatively small ensembles produce good results, and that it may be wasteful to use larger ensembles.

Our 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-143458. 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

1. C. Kamath and R. Musick, "Scalable data mining through fine-grained parallelism: The present and the future," in *Advances in Distributed and Parallel Knowledge Discovery*, H. Kargupta and P. Chan, eds., pp. 29–77, AAAI Press/The MIT Press, 2000.
2. C. Kamath, E. Cantú-Paz, I. Fodor, and N. Tang, "Searching for bent-double galaxies in the first survey," in *Data Mining for Scientific and Engineering Applications*, R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu, eds., Kluwer, Boston, MA, 2001.
3. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, CRC Press, 1984.
4. J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.
5. R. H. Becker, R. White, and D. Helfand, "The FIRST survey: Faint images of the radio sky at twenty-cm," *Astrophysical Journal* **450**, p. 559, 1995.
6. "FIRST: Faint images of the radio sky at twenty centimeters." http://sundog.stsci.edu/.
7. R. L. White, R. Becker, D. Helfand, and M. Gregg, "A catalog of 1.4 GHz radio sources from the FIRST survey," *Astrophysical Journal* **475**, p. 479, 1997.
8. R. L. White, 1999. Private Communication.
9. 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*, vol. 33, April 2000.
10. 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.
11. T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning* **40**(2), pp. 139–158, 2000.
12. E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning* **36**(1/2), pp. 105–139, 1999.
13. D. Opitz and R. Maclin, "Popular ensemble methods: an empirical study," *Journal of Artificial Intelligence Research* **11**, pp. 169–198, 1999.
14. L. Breiman, "Bagging predictors," *Machine Learning* **26**(2), pp. 123–140, 1996.
15. Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, L. Saitta, ed., pp. 148–156, Morgan Kaufmann, (San Mateo, CA), 1996.
16. L. Breiman, "Arcing classifiers," *Annals of Statistics* **26**, pp. 801–824, 1998.