Exploring Data Mining Techniques on Toy Data Set

Anna L. Villani, Charles Lanza, Casey Dawson
Department of Mathematics and Computer Science, College of the Holy Cross, Worcester, MA 01610, USA

This paper exhibits the findings of a exploratory project of classification a toy data set using various different data mining techniques. A large dataset was analyzed, pruned, and run through various data mining algorithms to determine some sort of classification. The goal of this project is to understand how manipulating a dataset can affect classification, understand different classification algorithms and how they can perform better/worse on particular datasets, and determine why this kind of work is important in real world applications.

I. The Data

The data set used is called zoo.arff; it was created by Richard Forsyth in May 1990. It contains data involving a variety of different types of animals. There are 101 instances, and 18 attributes in the file. There is no missing data. All of these attributes have Boolean values, except for the legs attribute which has a numeric value. One of these attributes, the animal type, is used as the class. These attributes are hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, and catsize. There are six possibilities for the class. These are mammal, bird, reptile, fish, amphibian, insect, and invertebrate. All living things are classified into a subgroup called class. In the animal kingdom, the classes under the phyla vertebrate and invertebrate are the exact classes described in this dataset. There are three other phyla belonging to the animal kingdom, but they are not represented in this dataset.

II. Pruning the data

One problem that was encountered with this data was that it was very small. This was solved by creating a test set from scratch, by researching other animals not included in the original data. The class values of this new dataset were evenly distributed and the attribute values were determined carefully, so that it corresponded correctly with the original instances. This solution was beneficial because it lengthened the overall data set and allowed that original set to be used purely for training. Another issue that was observed was the effectiveness of all the attributes. After testing the data on a variety of different algorithms, it became clear that not all of the attributes had an effect on classifying the data. For example, the milk and eggs were both attributes in the original data set. However, they were both used for the same purpose, to determine if the animal is a mammal or not. It was evident that the eggs attribute was superfluous, so it was removed. Another attribute that was deemed unnecessary was catsize. This attribute had a very wide range; from the size of a house cat to the size of a mountain lion. This attribute did not seem to be useful in any of the classification, so it was also removed. A third issue involved the instances in the original data set. One of the instances was a vampire. This seemed inappropriate because vampires do not exist. Its removal showed a slight improvement in the results, and also kept the data realistic. Another instance, frog, was in the data set twice. From observing these two instances, it was determined from the venomous attributes that one of the frogs was poisonous. Since these two instances are realistic and did not worsen the results, hey were both kept in the set.

III. Problem Description

The data mining problem solved for this data set was classification. Based on the 16 attributes describing each instance, the machine learning algorithms decided what type of animal each instance was. The algorithms were trained on the original 100 instances. Then it was determined how well these algorithms could correctly classify the 23 new instances that were created and used as the test data. Two different algorithms were chosen at the start of the experiment to conclude which knowledge representation model would work best with this data set. NaiveBayes is an algorithm that assumes that all of the attributes are independent and calculates the probability an instance belongs to a given class. This algorithm was chosen because after the data was cleaned and some attributes were removed, it seemed that the rest of the attributes should be used in the classification. Another algorithm that was chosen was the J48 decision tree. This was chosen based off knowledge of a simpler version of this algorithm called ID3. The decision tree works well for this data set because it still considers all of the attributes in the classification, but it recognizes that some attributes classify better than others. The J48 algorithm creates a tree by deciding what attribute to split on by calculating the information gain of each attribute. The information gain shows how much knowledge each attribute generates for the instances. For example, when this algorithm was run on the data set, the first attribute chosen to split was feathers. When the value was true, it could immediately conclude that the animal type was bird. The next attribute the algorithm chose to split on was milk. If the value of milk was true, it immediately concluded the animal type was mammal. Thus, J48 makes the smallest tree possible by choosing the attributes that classify the most instances. After testing these two algorithms on the data, it was determined that the J48 algorithm generated more accurate results. Also, the tree shows how the instances are classified in a clear and simple way. From this preliminary test, it was confirmed that trees would be the most effective way of representing the knowledge. Other than J48, various different tree algorithms were considered. One that was chosen was the NaiveBayes Tree. This algorithm is a combination of a decision tree and the NaiveBayes classifier. It is a two-level classifier tree with a decision tree classifier as the root node and several NaiveBayes classifiers as leaf nodes. This algorithm seemed appropriate because it utilizes the

advantages of both these algorithms, and both of them were considered during the process of choosing a knowledge representation model. Also, NB trees generally outperform decision trees and NaiveBayes alone. NB trees are good for datasets where many attributes are relevant, the attributes are not necessarily independent, the database is large, and interpretability of the classifier is important. Most of these situations apply to the zoo dataset used in this project, which further supports the choice of this algorithm.

IV. RESULTS

The 23 instances that were created were added to the original 100 instances to form a file containing a total of 123 instances. An 81.5% percentage split was chosen for this data set so that the original 100 instances were used to train, and the new ones were used to test. The results for the NaiveBayes tree are represented in **Table 1**. The confusion matrix showed that this classifier misclassified one reptile as an amphibian, one insect as an invertebrate, and one invertebrate as an insect. The kappa statistic is very close to one, so it is clear that it is performing better than random classification. The tree representation is shown in **Figure 1.** This tree shows that at each root node, the decision tree is playing its part by choosing which attribute to split on. At the leaf node, NaiveBayes takes over to finish the classification. Each leaf node represents a different NaiveBayes classifier. The results for the J48 algorithm are shown in **Table 2.** The confusion matrix for this algorithm shows that it misclassified one reptile as an amphibian and one insect as an invertebrate. Again, the kappa statistic is close to one so it is better than random classification. The tree representing these results is shown in **Figure 2.**

V. ANALYSIS

The J48 tree and NaiveBayes tree had very similar classification errors. They both misclassified an alligator, which is a reptile, as an amphibian. This likely happened because alligators were described as aquatic animals. The main difference between amphibians and reptiles is that amphibians live in water and on land, but reptiles can only survive on land. Although alligators spend much of their time in the water, they cannot ever breathe underwater. The description of them as aquatic probably led the classifiers to believe that it was an amphibian. Both of the classifiers also misclassified an earwig as an invertebrate. Since all insects are invertebrates, their attribute values are very similar. Thus, it is reasonable that the classifiers would mix them up. The NaiveBayes tree also misclassified a centipede as an insect, but the J48 tree did not. J48 ended up being a better classifier than NB tree, which is somewhat surprising because NB trees are said to generally be better than a decision tree alone. A possible reason for this is that NB trees work well for large datasets. In this case, the dataset may have been too small for the benefits of combining NaiveBayes and decision trees to be recognized.

VI. APPLICATIONS

There are a couple good real-life applications involving classification of animals. One is when zoos are being built, the animals need to be classified into distinct groups so that they can be placed in the correct habitats and receive proper care. For example, amphibians live on land and in water, so they need an environment that contains both of these things. Another example is birds. Most of them need an environment where they have room to fly. A second application is discovery of new animals. Although it is rare, it is possible for a new species to be discovered. In this case, its class would need to be determined.

VII. Figures

Correctly Classified Instances	20	86.9565%
Incorrectly Classified Instances	3	13.0435%
Kappa Statistic	0.847	
Total Number Instances	23	

Table 1.

Correctly Classified Instances	21	91.3043%
Incorrectly Classified Instances	2	8.6957%
Kappa Statistic	0.8978	
Total Number Instances	23	

Table 2.

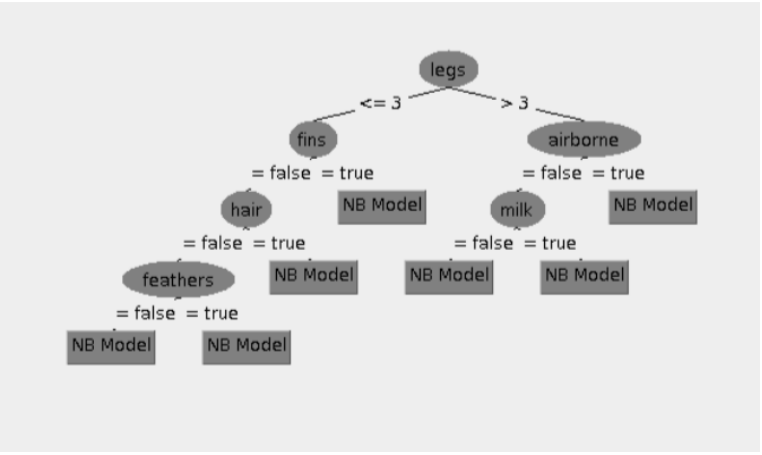

Figure 1.

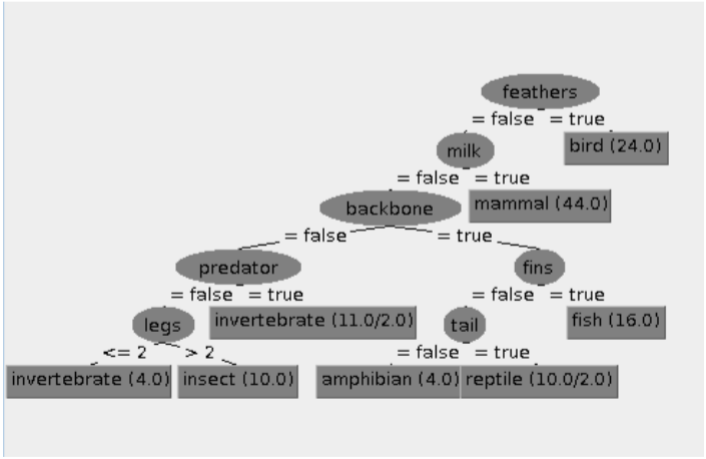

Figure 2.

REFERENCES

[1] "Animal Classification." - Reference. OpenCrypt Membership Software, n.d. Web. 14 Apr. 2014. <http://a-z-animals.com/reference/animal-classification/>.

[2] "English." Wikipedia. Wikimedia Foundation, n.d. Web. 14 Apr. 2014. <http://www.wikipedia.org/>

[3] Forsyth, Richard. "Index of /Datasets/UCI/arff." Index of /Datasets/UCI/arff. N.p., 15 May 1990. Web. 14 Apr. 2014. <http://repository.seasr.org/Datasets/UCI/arff/>.

[4] Kohavi, Ron. "Scaling Up the Accuracy of Naive Bayes Classifiers: A Decision-Tree Hybrid." Penn State. Silicon Graphics Inc, n.d. Web. 14 Apr. 2014. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.298&rep=rep1&type=pdf>..