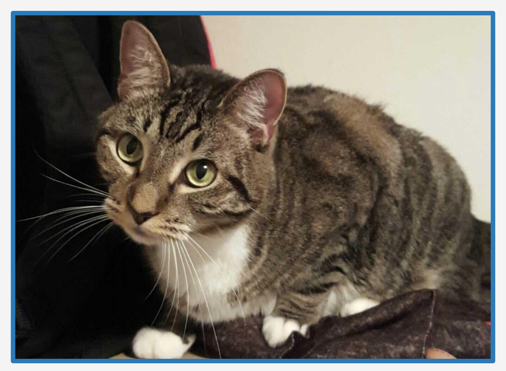
An Evaluation of Machine Learning Algorithms for Classification of Shelter Animal Outcomes

Mikayla Timm and Eman El-Sheikh, Ph.D.
Department of Computer Science
University of West Florida

University of West Florida
Computer Science

Introduction

Background



- Every year, **animal shelters** across the U.S. give approximately **7.6 million animals** a chance at finding a forever home and starting a new life [1].
- **Certain attributes** of the animals recorded at shelter intake may **affect their outcomes**.
- Shelters can **focus their efforts** on animals who are less likely to be adopted if they know the animal is at risk of having a **negative outcome**.

Goals

- Develop a system utilizing **machine learning** algorithms to **predict the outcomes** of animals brought to shelters based on the attributes recorded about them.

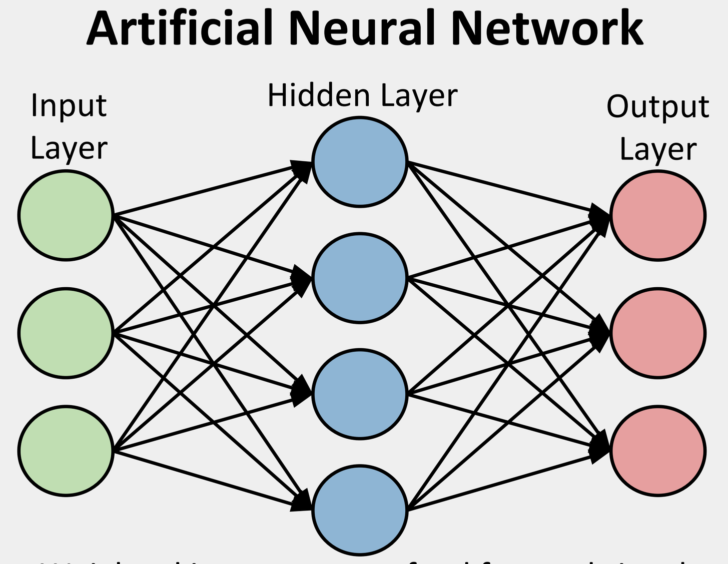
Approach

Data

- Data taken from Austin Animal Shelter from October, 2013, to March, 2016.
- **Attributes:** Name, Animal Type, Sex, Intactness, Age, Breed, Color, Date/Time
- **Possible Outcomes:** Adoption, Return to Owner, Transfer, Euthanasia, Died
- **Number of Animals in Training Set (outcomes recorded):** 26729
- **Number of Animals in Testing Set (outcomes not recorded):** 11456

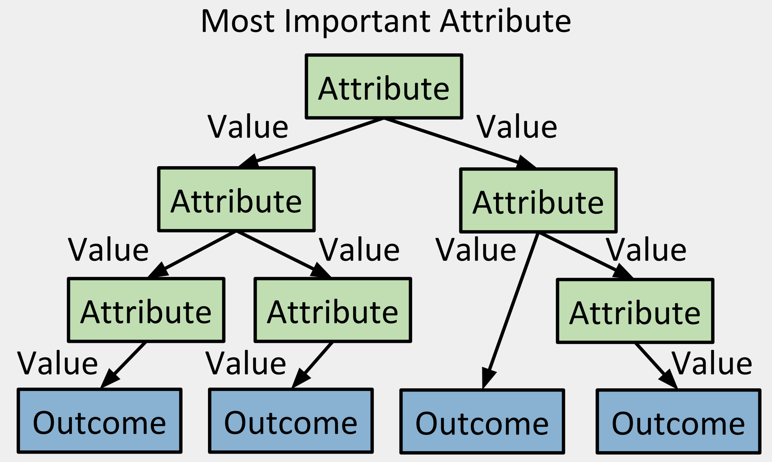
Classification

Artificial Neural Network



- Weighted input **neurons** feed forward signals through **axons** when input weights surpass a threshold in an **activation function** [2], [3]
- **Output layer** represents all possible **outcomes**
- **Many parameters** can be altered, including learning rate, momentum, number of epochs, and number of hidden layers and neurons
- **Relatively slow** - large datasets require a much longer training time for accurate classification
- More **complex** structure and algorithm

Decision Tree Learning

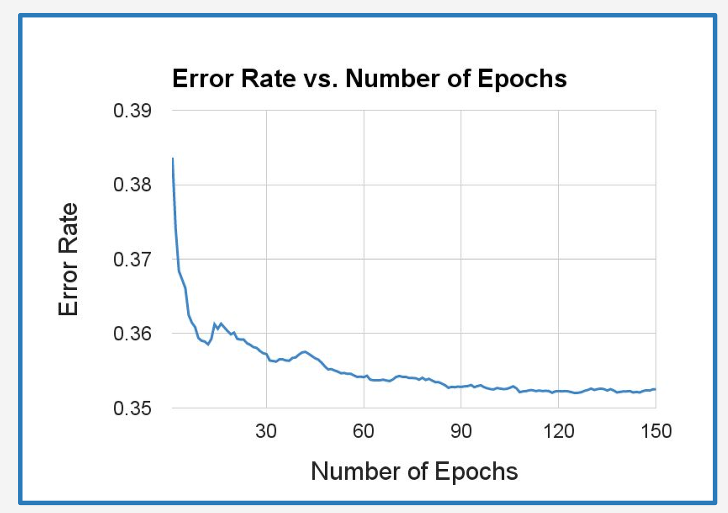


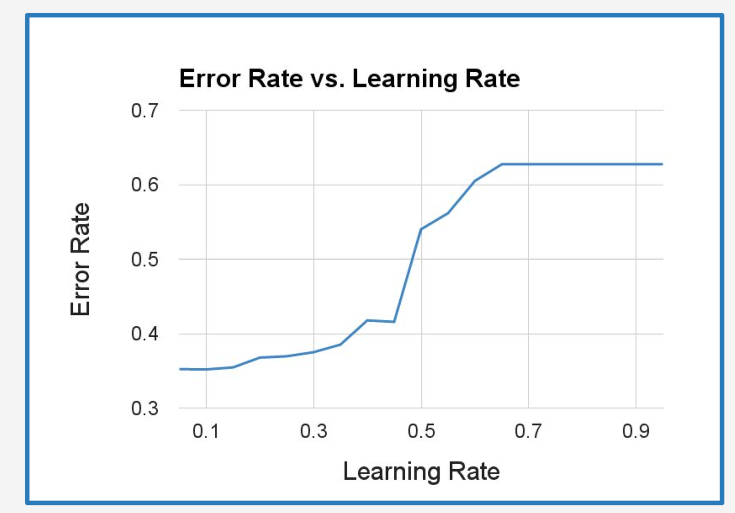
- **Split instances** by the values of the **most important attribute**, creating **subtrees** that continue to split until all instances in each leaf have the **same classification** [2], [3]
- Most important attribute determined by measuring **information gain** for each attribute - **difference in entropy** between parent and child
- **Relatively fast** - can sometimes classify accurately without looking at all attributes
- **Simpler** structure and algorithm

Results

Artificial Neural Network Parameterization

- By altering the **learning rate** and **number of epochs** for the neural network, we can observe changes in the **total error** for classification of the training set.



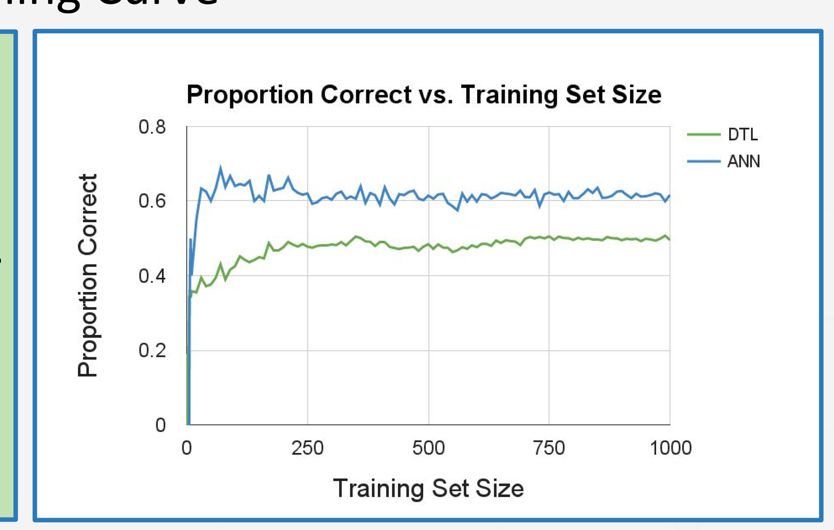


- Increasing the **number of epochs** leads to a **lower error rate** for the ANN.
- Increasing the **learning rate** to a point leads to a **higher error rate** for the ANN.

Artificial Neural Network (ANN) vs. Decision Tree Learning (DTL)

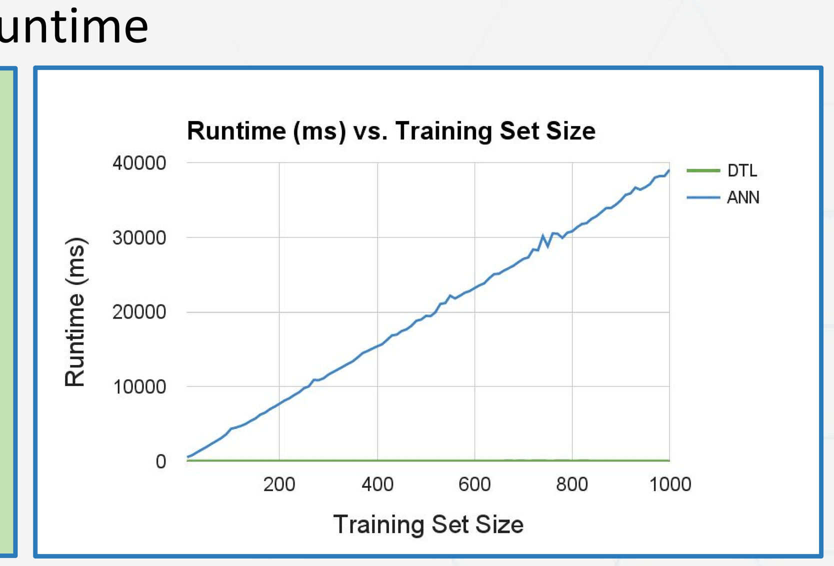
Learning Curve

- Comparative **learning curves** show the proportion of the training set **classified correctly** by each algorithm as the **training set grows**.
- From the graph, it appears that the **ANN** outperforms the DTL algorithm in terms of **accuracy**. The DTL grows at a slower initial rate.



Runtime

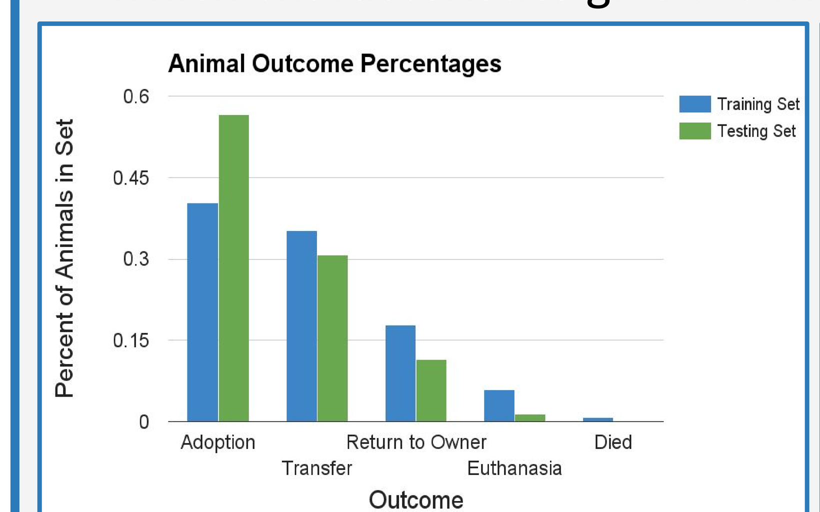
- This graph compares the **runtime** (in milliseconds) of each algorithm as the **training set size increases**.
- The **ANN's** runtime **increases** at a **linear** rate as the dataset grows.
- The **DTL's** runtime appears to not increase at all, staying **close to 0ms** even as the dataset grows large.

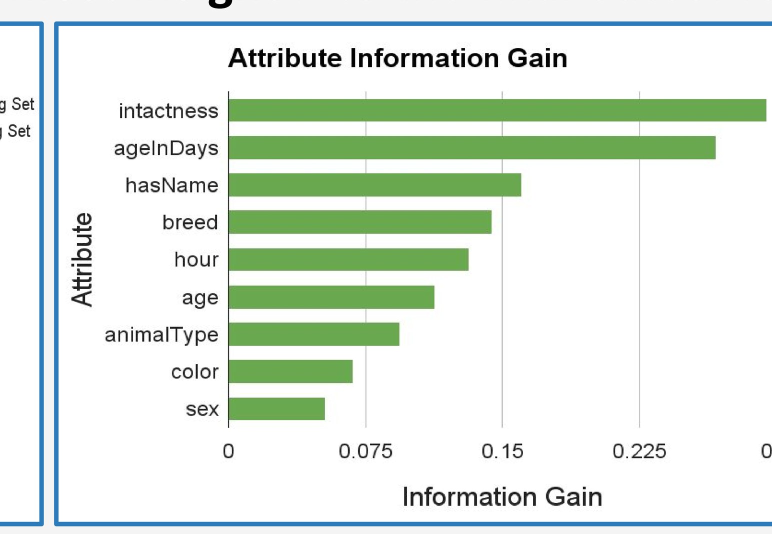


- It is apparent that there is a **trade-off** between algorithms' **accuracy** and **speed**.
- While the **ANN** takes a **longer time to run** on larger datasets compared to the DTL, it also reaches a relatively **high accuracy** given few examples. The **DTL** is significantly **faster**, but requires many **more examples** to reach a high accuracy.
- The **ANN** also has a **more complex** implementation than the **DTL** algorithm.

Results (cont.)

- The percentage of each outcome **predicted** in the **testing** set is relatively **similar** to the percentages found in the **training** set.
- By ranking the attributes by their information gain, we can see **which characteristics** give the **most insight** into the final outcome.





Conclusion

Discussion

- The **ANN** classifies a **higher percentage of correct instances** than the DTL, but takes a **much longer time** to run (no free lunch).
- Some **attributes** of the animals **do not matter much** (sex, color), while some make a **big difference** (intactness, ageInDays) in determining the outcomes.
- Given a greater number of **examples**, both algorithms **improve** in **accuracy**, though the runtime increases significantly for the ANN.
- Some attributes cannot be controlled, but **adoption rates** may **increase** by **spaying/neutering** the shelter animals and giving each one a **name**.

Future Directions

- Add decision tree **pruning** to combat overfitting issues
- Implement **GUI** for viewing visualizations of attributes/outcomes
- Test **other algorithms**, such as Naive Bayes, and compare results

References

1. **ASPCA**. Pet Statistics. http://www.aspca.org/animal-homelessness, 2016.
2. **M. Aly.** Survey on Multiclass Classification Methods. 2005.
3. **S. Russell and P. Norvig.** Artificial Intelligence A Modern Approach, Third Edition. 2010.

Acknowledgements

This project is inspired by the Kaggle competition "Shelter Animal Outcomes" on www.kaggle.com. The input data is provided by Austin Animal Center.
The Neural Network framework (Multilayer Perceptron) is provided by Weka.

UNIVERSITY of WEST FLORIDA