Using Decision Tree Learning as an Interpretable Model for Predicting Candidate Guide Dog Success

Zach Cleghern[1], Margaret Gruen[2], David Roberts[1]

1 Department of Computer Science, North Carolina State University; 2 College of Veterinary Medicine, North Carolina State University

Training guide dogs is a time and resource intensive process that requires a copious amount of skilled professional and volunteer labor. Even among the best programs, many dogs are released from their training programs. The highest cost in producing guide dogs occurs during the professional training and placement. Selecting dogs for training is a crucial task and guide dog schools can benefit from both an increase in accuracy of their selection and the speed at which dogs can be screened out of the program. We present a method using decision trees to predict the future success or failure of dogs in a guide dog program based on existing data sources from a guide dog school. We achieved 60.6% accuracy on predictions in the test set compared to a failure rate of about 50% at this stage in dog training. Decision trees are easily interpretable; thus evaluators can benefit not only from the model's prediction but can also examine which features are used to determine success or failure.

Introduction

The process of training dogs as service animals is highly resource intensive. The monetary and time demands vary by type of work, but some estimates show the cost for an organization to raise and train a service dog may be $20,000 to $50,000 per dog[1]. Unfortunately, many dogs that enter training do not succeed as guide dogs, representing a major loss to organizations. Great effort has been put forward toward attempts to identify successful dogs early in life, allowing for enhanced selection and refinement of breeding. The majority of these have focused on behavioral and health traits. However, several studies have also evaluated maternal care and early life experiences[2][3]. Evaluation and screening takes place at several points including neonatal, puppy-raising and training phases, including assessments as puppies, adolescent, end of training, and results following placement. At Guiding Eyes, observations from most assessments are scored using the Behavior Checklist. While some associations with success have been found, particularly at older ages, the predictive value of puppy testing has remained relatively low because environment plays a key role.

Guiding Eyes for the Blind (Guiding Eyes), a large, not-for-profit guide dog organization in the United States, maintains records of results for these assessments from all their dogs. At Guiding Eyes, socialization begins at 1 week and abservations of socialization events are recorded. A formal puppy assessment is conducted at around 7.5 weeks of age. About 11% of puppies tested are provided to other organizations that train service dogs, 16% are adopted as pets and the remaining 73% are raised as potential guide dogs by volunteer puppy raisers, who raise them in their homes, teach basic obedience and social skills and provide exposure to a variety of environments and experiences. Puppies are seen weekly then twice monthly by staff who also conduct formal assesments at 4 and 10 months of age. Their next assessment, the In-For-Training (IFT) test, is performed at a Guiding Eyes school when the puppy is about 16 months old. Following the IFT test, dogs may enter into training as a guide dog, enter training for other work (detection or emotional support), or be released as pets. This data corpus provides an opportunity to employ machine learning to assist in decisions about which dogs to invest training resources in.

Machine learning can identify subtle, otherwise difficult to discover patterns present in data. By incorporating data from both successful and unsuccessful dogs ("success" here means the dog was successfully placed with a handler and/or became a successful breeding dog), machine learning methods can detect patterns that are reflective of dog performance in ways that can be difficult for humans, even experts, to pick up on. Then, the results can be used to better understand early determinants of future success.

We chose decision trees due to their interpretability, which can be used to find features that are highly predictive of training outcome. The induction of decision trees[4] for classification is a standard method[5] in the realm of supervised machine learning. In this case the dependent variable, or "class", is a label indicating whether a dog is

R.A Grant et al. (Eds.): Measuring Behavior 2018, ISBN 978-1-910029-39-8
Manchester, UK, 5th-8th June 2018

successful in the program or not. We make predictions with decision trees by examining a set of *features*, or attributes, about a specific dog in the data set. A decision tree is a graph structure like a flowchart consisting of *nodes*, which represent a test on a condition, and connections between the nodes called *edges*, which represent the result of the test. By starting at the *root node*, or first decision, one can follow the edges to arrive at a *leaf node*, containing the prediction produced by the tree. The nodes between the root and leaf are *internal nodes*, which represent tests on specific attributes. For example, an integer attribute X and a node in the tree might test "X <= 0." When traversing the tree at this node, the expression is evaluated using the attribute X of the current data sample (e.g. the data associated with a single dog) and the edge corresponding to the result is selected.

In machine learning, an algorithm, such as CART[6] (Classification and Regression Trees- what we used), constructs decision trees by analyzing a training set of data samples. Induction of the tree happens incrementally beginning with the root node. Using some measure of impurity or information gain, such as Gini Impurity[7], the features are analyzed for the best possible split according to the measure. A feature and threshold value are chosen that splits the trainind data accordingly. Then each split is examined, creating new nodes until a stopping condition is met, such as maximum tree depth. Maximum tree depth is a number specified outside of the learning process, a *hyperparameter*. Other hyperparameters in decision tree learning are splitting criterion (how to measure each feature to select the next split), minimum samples in a split to add a new node, and maximum number of features to consider for a single tree.

Using the induced tree, we can classify a data sample by starting at the root node and testing the features at each node, following the associated path down the tree until a leaf node is reached. Once this happens, we have our prediction of the sample's class. Because each node explicitly states the attribute and threshold used in the test, this process is easily interpretable by humans allowing non-experts to use decision trees without computer interaction and also understand which features were relevant in the decision. By viewing the nodes, we can observe which features provide the most information about how to classify any sample. See Figure 1 for an example portion of a decision tree with 3 nodes shown. The first feature tested is an item from the Behavior Checklist involving the dog's confidence; the number of samples indicates how many data samples were in the portion of the data that this node is concerned with.

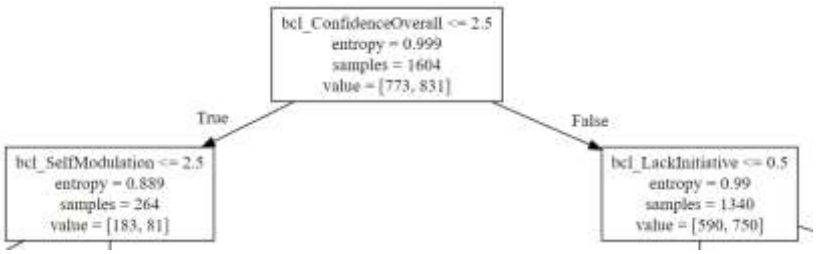

Figure 1. Example portion of decision tree with 3 nodes, each using Gini Impurity[7] splitting criterion.

The goal of this study is to show how machine learning, using existing behavioral data, can be employed to predict which dogs are more likely to succeed as guide dogs. Our objectives were to use decision tree learning with existing data collected by Guiding Eyes, including the Behavioral Checklist and the CBARQ test, and then to test the developed model using an unseen data set. If successful, our model would allow early predictors of success to be identified and used to increase accuracy of early career sorting and identifying which dogs are most appropriate for training. Ultimately, this would benefit the guide dog schools, the people they assist, and the dogs themselves. It is worth noting that our efforts are not the first to specifically address the problem of predicting success of guide dogs with advanced analytics and machine learning. Yim et al.[8] also used decision trees to predict guide dog success, but we are extending and improving on this work by restricting the time period to enable earlier predictions than previously available; ultimately this allows guide dog schools to filter out candidate dogs earlier in life and conserve valuable resources.

Data

Data were available for 1561 Labrador retrievers (765 males; 796 females) produced by Guiding Eyes over several years. Of these, 789 (50.5%) were successful in the guide dog program. Our definition of "success" here means that the dog was successfully placed with a person who is blind or visually impaired or possessed exceptional qualities and was selected as a breeding dog. The data used for the machine learning task included basic information about the dog (sex and breed) and the results of two common forms of evaluations, described here.

CBARQ

The CBARQ[9] is a 101 item questionnaire used to evaluate a dog's behavior and temperament. This questionnaire was developed by Serpell and Hsu, and has been validated for use in dogs[10]. It is commonly used in canine research [11][12], and comprises data that many guide dogs schools, in addition to Guiding Eyes, already collect[13]. The data we used from this questionnaire was collected when dogs were 12 months of age.

Behavior Checklist

Also developed by Serpell, the Behavior Checklist is a scoring system where observers can score up to 52 aspects of behavior such as excitability, fearful behavior (such as traffic or noise fear), emotional response to potentially stressful situations, and body sensitivity. Like the CBARQ questionnaire, the Behavior Checklist is a commonly used scoring system for scoring the behavior of potential guide dogs, which makes the test a good source of data if we want models that are useful for any guide dog school. The Behavior Checklist we utilized was collected at the In-For-Training evaluation.

Combined with the basic information about the dog (sex and a categorical breed code), these information sources result in 197 individual features. These features were tested for correlation with outcome in order to derive a restricted set of significant features for testing. Twenty-seven features had statistically significant ($p < 0.05$) correlation with success. These significant correlations in the 12 month CBARQ and Behavior Checklist data are shown in Tables 1 and 2, respectively. In the CBARQ questionnaire[9][10], the questions referred to in Table 1 are as follows:

4. Q10: [Displays aggressive behavior] when approached directly by an unfamiliar adult while being walked/exercised on a leash.
5. Q11: [Displays aggressive behavior] when approached directly by an unfamiliar child while being walked/exercised on a leash.
6. Q12: [Displays aggressive behavior] toward unfamiliar persons approaching the dog while s/he is in your car (at the gas station, for example).
7. Q52: [Displays fearful behavior] when having his/her feet toweled by a member of the household.
8. Q85: Nervous or frightened on stairs.

Table 1. CBARQ features in the data set with a statistically significant correlation with guide dog success.

Feature	Correlation	P
CBarq-Q10	-0.056	0.0269
CBarq-Q11	-0.063	0.0133
CBarq-Q12	-0.054	0.0341
CBarq-Q52	-0.05	0.0478
CBarq-Q85	-0.063	0.0131

Table 2. Behavior Checklist features in the data set with a statistically significant correlation with guide dog success.

Feature Name	Correlation	P Value	Feature Name	Correlation	P Value
BCL-	-0.052	0.0383	BCL-DogAggression	-0.196	$5.15*10^{-}$
BCL-FearDogs	0.053	0.0376	BCL-	-0.196	$5.96*10^{-}$
BCL-	0.118	$3.14*10^{-}$	BCL-	-0.14	$2.66*10^{-}$

BCL-	-0.051	0.0429	BCL-	0.051	0.0429
BCL-Footfall	-0.052	0.04	BCL-HandlerDogTeam	-0.263	$3.97*10^{-}$
BCL-	-0.147	$5.24*10^{-}$	BCL-PressuresDog	-0.185	$1.69*10^{-}$
BCL-	-0.177	$1.99*10^{-}$	BCL-KennelsPoorly	-0.055	0.0286
BCL-	0.052	0.0419	BCL-KennelAnxiety	-0.195	$8.94*10^{-}$
BCL-Olfactory	0.141	$2.32*10^{-}$	BCL-ChewsValuables	-0.184	$2.54*10^{-}$
BCL-Scavenges	0.17	$1.56*10^{-}$	BCL-Housebreaking	0.204	$4.45*10^{-}$
BCL-	-0.177	$1.97*10^{-}$	ComparisonRating	-0.057	0.0245

Building the Decision Tree Model

Decision trees are easy to understand by humans, which make them a suitable model for our prediction problem in which professionals evaluating dogs will want to know why a model chooses a particular prediction. We used the Python library scikit-learn[14], an easy to use machine learning package, specifically the CART algorithm[6]. We did not require much preprocessing of our data aside from reformatting some information to be usable by the scikit-learn library. For efficiency of the model building process, we removed any features (such as dog breed, since only Labrador Retriever data was submitted) whose values were equal for every dog in the training set. To evaluate the quality of the learned models, we divided the data into training and test sets, with a split of 80% in the training set and 20% in the test set, which was only used once the hyperparameters were optimized and we had trained the final model.

The hyperparameters to the induction algorithm that we optimized were maximum depth, splitting criterion (in scikit-learn, either the Gini Impurity[7] or Information Gain[7]), and minimum number of samples on which to split an internal node and also on a leaf node. To find the best set of values, we used grid search[15]. This is a simple method for finding sets of hyperparameter values that exhaustively searches the combinations of values in a search space by applying the learning algorithm and comparing the results on validation sets. Although more optimal approaches exist[15], grid search was sufficient for our small search space and the process did not take more than 5 minutes for even the full feature set. For each set of possible values, the algorithm further divides the training set into training and validation sets using 3-fold cross validation[16], which calculates a mean accuracy over the validation sets. The hyperparameters were chosen from the best scoring mean accuracies. We then used those values to fit a tree model on the entire training set. Finally, we evaluated the quality of the resulting decision tree on the previously unused test set using three metrics: accuracy, precision, and recall. (we did this for several subsets of the full feature set). *Precision* is the fraction of true positives (successful dogs that were also predicted to be successful) divided by the true and false positives (the number of times the model said the dog would be successful). *Recall* is the fraction of true positives divided by the number of successful dogs in the test set.

Results

The best set of hyperparameters for the full feature set were found to be: maximum depth of 5, entropy (information gain) splitting criterion, 6 as the minimum number of samples for a leaf node and 9 as the minimum samples for an internal node. We also tested the effectiveness of just the set of statistically significant features previously shown in Tables 1 and 2.

Table 3. Best cross-validation scores and test set accuracy, precision, and recall for each feature subset. Note that "breed" was actually removed as every dog in the data set had the same breed code.

Feature Set	Cross-validated Mean	Test Set Accuracy	Test Set Precision	Test Set Recall
Basic Dog Info (sex, breed)	0.5152	0.4856	0.5161	0.4819
BCL	0.638	0.608	0.62	0.698

CBarq	0.5521	0.5144	0.5507	0.4578
Info, BCL, CBARQ	0.635	0.606	0.620	0.684
Statistically significant	0.648	0.596	0.698	0.698

Table 3 shows the metrics obtained for several subsets of features. Allowing the learning algorithm to choose only from the set of features across all data sources whose correlation with success was statistically significant ($p < 0.05$) resulted in only slightly worse accuracy than the full feature set, but actually outperformed in terms of precision and recall. The full feature set achieved a 60.6% accuracy on unseen data, but since dogs at this point in time have a 50% failure rate this shows promise for the decision tree learning approach. Several subsets achieved recall scores of almost 70%. This indicates the decision tree models were effective at identifying the successful dogs (though false positives brought down the overall accuracy).

By using a decision tree *regressor* instead of a classifier, we were also able to learn decision trees that output a number between 0 and 1 and can be interpreted as a probabilistic prediction. To evaluate the quality of these predictions, we calculated the mean absolute value of the error, where *error* is the difference between the predicted output and the actual class. We treated success as the real number 1.0 and failure as 0. The results of this are shown in Table 4.

Table 4. The mean absolute error for each feature set when converting the prediction task to a regression problem.

Feature Set	Mean Absolute Error
Basic Dog Info (sex,	0.500
BCL	0.434
CBarq	0.488
Info, BCL, CBARQ	0.472
Statistically significant	0.464

Discussion

Using decision tree learning, we were able to successfully predict the future success or failure of potential guide dogs with 60.8% accuracy, 62% precision, and 68.4% recall on the full feature set. Since the eventual failure rate of dogs at this time point (the IFT test) is close to 50%, this is suggestive of an opportunity for cost savings for guide dog programs, especially if this accuracy score can be improved upon. Incorporating larger data sets from various guide dog schools could improve this accuracy. By examining sources of behavioral data that already exist, decision tree classifiers can extract patterns for guide dog schools to utilize in evaluating dogs for training. Including only features with significant correlations to success code was much faster (about 4 minutes and 4 seconds for the full feature set; 1 minute and 9.5 seconds for the significant features), but this is unlikely to be an issue in this domain.

Understanding what influences the ability of dogs to work in various roles, behaviorally or otherwise, has been a significant and often multidisciplinary area of research. There are studies that analyze what factors can tell us about the future behavior of dogs, such as the work by Weiss[17] in which behavioral tests were used to aid in the selection of shelter dogs of service dog training. The behavioral tests include assessments of reaction to touch, walking, fetching, and interacting with strangers. They then used regression analysis to predict the dog's performance over 5 weeks of training. Slabbert and Odendaal found in a two year longitudinal study several behavioral tests whose correlation with future success as police dogs was statistically significant. Some of these tests were able to be administered as young as 8 weeks of age[18]. Byrne et al.[19] used instrumented dog toys with sensors to develop a logistic model tree for classifying the eventual outcome of dogs in a service dog organization (Canine Companions for Independence) and were able to achieve an accuracy of 87.5%, which by their calculations could save $70,000 by identifying dogs that will likely fail the program.

Arata, Momozawa, and Takeuchi[20] studied behavioral factors that may predict success or failure in guide dog programs with the use of a questionnaire. They were able to assess that distraction, sensitivity, and docility are

important behavioral factors that can predict a dog's future success. They found that distraction could predict a dog's success with 80.6% accuracy. Batt et al.[21] evaluated the ability of several behavioral tests to predict guide dog success and found several behavioral tests and physical characteristics that were predictive of success. Their results obtained at 14 months of age were more accurate than their results obtained at either 6 months or when the dog completed training. Using the CBARQ questionnaire, Duffy and Serpell found that tests administered at 6 and 12 months of age were useful for predicting success in guide and service dog programs, but their predictive value varied across organization[13].

A benefit of our approach is that we used data sets that are already commonly collected and did not implement our own tests. Our method is also able to provide a "line of reasoning" for its predictions by producing a path from root node to leaf node so that evaluators can know why a prediction was made. Probabilistic predictions may also prove useful by identifying dogs who may have high (but not exceptionally high) probability of success. This is useful for at least two reasons. Guide dog schools will be ultimately using their own discretion on the decision whether to train each dog and a probabilistic machine learning prediction may be more helpful in that regard. A second reason is these dogs may have specific issues that need to be addressed and their success could hinge on those issues.

Future Work

There are many avenues to explore in using machine learning to predict the future capabilities of guide dogs. Recent studies highlighting the importance of maternal care also open a new avenue for data that may refine our model[2][3]. Furthermore, Guiding Eyes collects a variety of text-based data associated with each dog at certain steps in its development such as training notes and comments recorded for various tests. Convolutional neural networks (CNNs) are well-suited to incorporating this data into a machine learning model alongside numeric features. Neural networks may be able to improve the accuracy but in contrast with decision trees can be very difficult to extract meaning from. In addition, IBM Watson[22] services for natural language processing may also be able to augment machine learning prediction of dog outcomes and also to tackle similar problems, such as optimal pairings of puppies with raisers.

Ethical Statement

This work utilized historical data provided by a not-for-profit organization and did not involve new data collection.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1329738. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was also supported in part by a faculty award from IBM. The authors would like to give an enthusiastic thanks to Guiding Eyes for the Blind for their support as well. The views expressed in this paper are the authors' own, and do not necessarily reflect the views of IBM or Guiding Eyes.

References

1. Berns GS, Brooks AM, Spivak M, Levy K (2017). Functional MRI in awake dogs predicts suitability for assistance work. *Scientific Reports 2017*; **7**:43704. doi:10.1038/srep43704
2. Bray EE, Sammel MD, Cheney DL, Serpell JA, Seyfarth RM (2017). Characterizing early maternal style in a population of guide dogs. Frontiers in Psychology; 2017 doi: 10.3389/fpsyg.2017.00175
3. Bray EE, Sammel MD, Cheney DL, Serpell JA, Seyfarth RM. Effects of maternal investment, temperament, and cognition on guide dog success (2017). Proceedings of the National Academy of Science; doi:10.1073/pnas.1704303114.
4. Quinlan, J. Ross (1986). Induction of decision trees. *Machine learning* **1.1**: 81-106.

5. Weiss, Sholom M., and Ioannis Kapouleas (1990). An empirical comparison of pattern recognition, neural nets and machine learning classification methods. *Readings in machine learning*. 177-183.
6. Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software
7. Fayyad, Usama M., and Keki B. Irani (1992). The attribute selection problem in decision tree generation. *AAAI*.
8. Yim, Brendan, Su, Kun, Siahpolo, Soraya, Tseng, Chris (August 2, 2017). Guiding Eyes IBM Watson Project. *Personal correspondence with the authors*.
9. Segurson, Sheila A., James A. Serpell, and Benjamin L. Hart (2005). Evaluation of a behavioral assessment questionnaire for use in the characterization of behavioral problems of dogs relinquished to animal shelters. *Journal of the American Veterinary Medical Association* **227.11**: 1755-1761.
10. Hsu, Yuying, and James A. Serpell (2003). Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *Journal of the American Veterinary Medical Association* **223.9**: 1293-1300
11. Marshall-Pescini, Sarah, et al (2008). Does training make you smarter? The effects of training on dogs' performance (Canis familiaris) in a problem solving task. *Behavioural processes* **78.3**: 449-454.
12. Serpell, James A., and Yuying A. Hsu (2005). Effects of breed, sex, and neuter status on trainability in dogs. *Anthrozoös* **18.3**: 196-207.
13. Duffy, Deborah L., and James A. Serpell (2012). Predictive validity of a method for evaluating temperament in young guide and service dogs. *Applied Animal Behaviour Science* **138.1**:99-109.
14. Pedregosa, Fabian, et al (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12.Oct (2011)**: 2825-2830.
15. Bergstra, James, and Yoshua Bengio (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13.Feb**: 281-305.
16. Kohavi, Ron (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*. **Vol. 14. No. 2. 1995.**
17. Weiss, Emily (2002). Selecting shelter dogs for service dog training. *Journal of Applied Animal Welfare Science* **5.1**: 43-62.
18. Slabbert, J. M., and J. S. J. Odendaal (1999). Early prediction of adult police dog efficiency—a longitudinal study. *Applied Animal Behaviour Science* **64.4**: 269-288.
19. Byrne, Ceara, et al (2018). Predicting the Suitability of Service Animals Using Instrumented Dog Toys. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1.4**: 127.
20. Arata, Sayaka, et al (2010). Important behavioral traits for predicting guide dog qualification. *Journal of Veterinary Medical Science* **72.5**:539-545.
21. Batt, Lara S., et al (2008). Factors associated with success in guide dog training. *Journal of Veterinary Behavior: Clinical Applications and Research* **3.4**:143-151.
22. High, Rob (2012). The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks*.