Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN 2011)

Bias of Importance Measures for Multi-valued Attributes and Solutions *

Houtao Deng[1], George Runger[1], and Eugene Tuv[2]

[1] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
{hdeng3, george.runger}@asu.edu
[2] Intel Corporation, Chandler, AZ, USA
eugene.tuv@intel.com

Abstract. Attribute importance measures for supervised learning are important for improving both learning accuracy and interpretability. However, it is well-known there could be bias when the predictor attributes have different numbers of values. We propose two methods to solve the bias problem. One uses an out-of-bag sampling method called OOBForest and one, based on the new concept of a partial permutation test, is called pForest. The existing research has considered the bias problem only among irrelevant attributes and equally informative attributes, while we compare to existing methods in a situation where unequally informative attributes (with or without interactions) and irrelevant attributes co-exist. We observe that the existing methods are not always reliable for multi-valued predictors, while the proposed methods compare favorably in our experiments.

Keywords: attribute importance, feature selection, random forest, cardinality

1 Introduction

Attribute importance measures for supervised learning are important for improving both learning accuracy and interpretability. There are well known attribute importance measures such as information-based measures, chi-squared, and so forth. However, the bias problem for multi-valued attributes has been recognized for these methods. We refer to the number of distinct values of a attributes as its cardinality. [2] noted that attribute selection with Gini gain measure is biased in favor of those attributes with higher cardinality. [11] showed that there are biases in information-based measures adopted by decision tree inductions. [4] showed that attribute selection biases not only exist in information gain measures such as the Gini index, but also in others such as the distance measure in Relief [5], etc.

For solving the multi-valued problem, [7] introduced a normalization into the attribute selection measure known as the gain ratio. However, attributes with

* This research was partially supported by ONR grant N00014-09-1-0656

very low information values then appeared to receive an unfair advantage [11, 4]. Also [11] experimented with discrete, uniformly distributed attributes with different number of levels. They concluded that chi-squared could be used for the multi-valued problem. [4] proposed a minimum description length principle to alleviate the feature selection bias, but also mentioned that there are still slight decreases in the importance measure with the increasing cardinality.

Recently, a conditional inference framework [3] was proposed to solve the overfitting and attribute selection bias problems. [9] showed that this method (referred to as cForest) demonstrated promising results in both null and power cases. In the null case, all predictor attributes are irrelevant with different cardinality. In the power case, only one predictor attribute is informative, all other attributes are irrelevant with different cardinality [9]. Though these research methods have successfully discovered and alleviated the multi-valued problem to some degree, we observe that there are still some important problems that are unresolved.

A permutation importance measure (PIMP) was introduced by [1]. It permutes the target attribute and a p-value can be used to measure importance. However, PIMP fits the importance score with a prior probability distribution. Though specifying a prior distribution is not necessary, [1] used prior probability distributions in their experiments. One of our proposed algorithms (pForest) also uses permutation importance, but there are substantial differences. pForest permutes the predictors, and more importantly, make use of a partial permutation strategy for better efficiency. Furthermore, pForest does not need prior probability distributions to be specified. Here we focus on non-parametric methods and thus compare our methods to cForest in the later experiments.

Most experiments from the existing research are limited to some idealized situations. For example, [11] considered the multi-valued problem only for irrelevant attributes, while [9] considered irrelevant attributes and only one informative attribute. [4] considered irrelevant and equally informative attributes. However, there can exist both irrelevant and unequally informative attributes with different cardinalities. Furthermore, the informative attributes may interact with each other.Therefore, it is important to consider the multi-valued problem under more realistic scenarios.

We propose two new solutions for these problems. We focus on two-class classification (a common supervised problem), but our methods can be extended. We also focus on tree-based ensembles because of their capability to generate robust models that can handle nonlinearities, interactions, mixed (categorical and numerical) attributes, missing values, attribute scale differences, etc. However, our second method is not limited to a certain type of classifier. It is a meta approach that can be applied to improve feature selection algorithms. Furthermore, we contribute a more comprehensive simulation framework for studying the problem that integrates multiple cardinalities, and where non-equally informative attributes (with or without interactions) and irrelevant attributes co-exist. Such a framework can provide a useful benchmark to compare alternatives. Section 2 briefly summarizes some widely used importance measures. Section 3 describes

our proposed attribute importance methods. Section 4 describes our simulation framework and our experimental results, while Section 5 provides conclusions.

2 Attribute importance measures

Random forest (RF) [6] is a commonly-used feature selection tool. It allows for not only nonlinear models, but also attribute interactions. However, it can suffer from the multi-valued problem because it is based on an information criteria. Consequently, a remedy for RF's problem is important.

RF builds an ensemble of decision trees. Each tree is built on a bootstrap sample (random, with replacement) from the original training data. Also, at each node only a subset of attributes is selected from the full set of attributes and the split is calculated only from members of this subset. The objective is to decrease the correlation between trees in the ensemble in order to decrease the final model variance. RF uses the Gini impurity criterion for scoring attribute importance. Denote $Imp(X_k, \tau)$ as the importance of a attribute X_i at a single tree τ, then $Imp(X_k, \tau) = \sum_{t \in \tau} \triangle Gini(X_k, t)$ where $\triangle Gini(X_k, t)$ is the Gini impurity decrease at a node t where X_k is the splitting attribute. The Gini index at node t is defined as $Gini(t) = \sum_j p_j^t(1 - p_j^t)$ where p_j^t is the proportions of cases of class j at node t. The importance of X_k is obtained from the sum of the importance scores from trees $\tau_m, m = 1, \ldots, M$ in a RF. For every tree τ in the ensemble, the instances not selected in the bootstrap sample are referred to as out of bag (OOB) and these cases can be considered to be a test sample for tree τ. These samples are used in our proposed importance measure.

A conditional inference framework [3] was proposed to solve the overfitting problem and attribute selection bias problem. [9] used the method to measure importance for multi-valued attributes in a model similar to a RF. In this method, for each node, first the attribute to be split is selected by minimizing the statistical p value of a conditional inference independence test. Then the splitting value is established by an appropriate splitting criterion. The separation of attribute selection and splitting criterion is the key to handle the cardinality bias [3].

3 Attribute importance from OOBForest and pForest

In this section we introduce new methods to score attribute importance. An OOBForest uses the training samples to find the best splitting value on each attribute in the same manner as for a RF (with the Gini index as the default information measure). But, instead of discarding the OOB samples when building a tree, the OOB samples are used to select the best splitting attribute at a node. That is, the Gini index is recomputed for the OOB samples based on the split value obtained from the training data at each node. Furthermore, the importance measure $Imp(X_i, t)$ uses only the OOB samples. The principle here is similar to a conditional inference framework. The attribute selection criterion and splitting criterion are separated. The role of OOB samples was discussed for model improvements in [10], here we propose to use it to specifically solve

the bias problem in measuring attribute importance. Computationally, the extra work over a RF is to calculate the split score from the OOB samples at each node in the forest. Because the OOB samples for a tree are typically smaller than the original training data less time is needed (approximately 2/3 less) than to generate a second RF (and the basic RF algorithm is fast [6]).

Next consider the pForest. Denote $X_k, k = 1, ..., K$ as the predictors and T as the target. [8] used permutation tests to obtain the statistical p value for dependency between an X_k and T. Then the inverse of the p value was used as the importance of the attribute. However, this method only measures the dependency of T over a single attribute X_k and the interactions between predictors are not considered. Permutation tests for feature selection was also used by [10]. Their method first randomly permutated each attribute $X_k, k = 1, ..., K$ and then compared importance score of an attribute to the distribution of scores from the irrelevant variables obtained from the permutations to obtain the corresponding attributes $Z_k, k = 1, ..., K$.

Our proposed algorithm also uses permutations, but an attribute is only compared to permuted version of itself. Furthermore, we introduce the concept of partial permutations. In each replicate r, by applying an importance method $f(.)$ (such as RF) to $\{X_k, Z_k, T, k = 1, ..., K\}$, the importance score of $X_k, Z_k, k = 1, ..., K$, that is, $Imp_r(X_k)$ and $Imp_r(Z_k)$ can be obtained. A feature X_k is compared directly to its permuted version Z_k in each replicate to match the cardinality between X_k and Z_k (and this differs from [10]). Next consider the measure

$$Imp(X_k) = \frac{1}{R} \sum_{r=1}^{R} I[Imp_r(X_k) > Imp_r(Z_k)] \tag{1}$$

where $I(.)$ denotes the indicator function. It can be seen that equation (1) is proportional to a binomial distribution $B(R, p_k)$, where p_k is the probability that $Imp(X_k) > Imp(Z_k)$. It is not feasible to compute the true p_k over all possible permutations in most practical situations. Therefore, [8] suggested a bounded number of permutations to achieve a significance level of 0.05.

The basic approach described so far is effective to distinguish informative from noninformative attributes. However, to rank informative attributes a more subtle refinement is used. In order to better detect finer importance relationships we propose partial permutations. That is, Z_k is obtained from permuting a fraction of the rows of X_k (a fraction δ selected randomly in each replicate). Consequently, as δ is decreased X_k and Z_k are more similar and it is more difficult for $Imp_r(X_k) > Imp_r(Z_k)$. Our default choice is $\delta = 20\%$ and in our experiments with the default δ, along with $R = 200$ replicates, we can achieve good results. We refer to this partial permutation method to attribute importance, with importance scores obtained from a RF, as the pForest.

Computationally, pForest is more demanding than OOBForest because each replicate requires another RF to be generated. However, the speed of a RF enables even hundreds of replicates to be computed in minutes for moderate data sets. Finally, note that although we focus on decision-tree ensembles, the permutation strategy to solve the multi-valued problem can be applied to any

feature selection method $f(.)$. One would simply replace the score $Imp_r(X_k)$ with another method and still average $I[Imp_r(X_k) > Imp_r(Z_k)]$ over the replicates.

Algorithm 1: pForest importance measure

Input: R = number of permutation replicates; δ = percentage of rows permuted;
training data $D = \{(x_i, t_i)|i = 1, \ldots, N\}$ with K features $F = \{X_1 ... X_K\}$
$f(F, D)$ is a function that provides importance scores for attributes in F with data D (default is RF)
for $r = 1, \ldots, R$ do
  $Z_k \leftarrow$ randomly select and permute $\delta * N$ rows of X_k, for $k = 1...K$
  set $F' \leftarrow F \cup \{Z_1, \ldots, Z_K\}$
  $Imp_r(F') = f(F', D)$
end for
$Imp(X_k) = \frac{1}{R} \sum_{r=1}^{R} I(Imp_r(X_k) > Imp_r(Z_k))$, for $k = 1...K$
Output: $Imp(X_k)$, for $k = 1...K$

---

4 Experiments

Similar to [11, 4, 9], the experiments are setup as simulations so that the "ground truths" for attribute importance are known. The relationships between the predictors and the target are shown in Figure 1. Here T_1 and T_2 are the target attributes with and without interactions present in the model, respectively. All other attributes are predictors. The generation and properties for these attributes are summarized as follows:

- Generate $X_1 \sim Normal(0, 10)$, and then discretize (equal-frequency) into X_2, X_3, X_4, X_5, X_6 with different cardinalities shown in Figure 1. Randomly permutate 30% of the rows of X_5, and 50% of the rows of X_6. This injects different amounts of noise into X_5, X_6 so that they are unequally informative concerning the target.
- Generate $Y_k, k = 1, ..., 6$ independent from $X_k, k = 1, ..., 6$. The generation procedure is similar to the generation of $X_k, k = 1, ..., 6$.
- Generate $U_1 \sim Uniform(-10, 10)$, and then discretize (equal-frequency) into U_2, U_3, U_4, U_5, U_6 with different cardinalities.
- Generate the binary target T_1 as $P(T_1 = X_4) = 0.95, P(T_1 \neq X_4) = 0.05$.
- Generate the binary target T_2 as $P(T_2 = XOR(X_4, Y_4)) = 0.95, P(T_2 \neq XOR(X_4, Y_4)) = 0.05$ (where XOR is the exclusive *or*).

Two experiments are derived from the relationships among the attributes. First T_1 is the target and $\{X_k, U_k, k = 1, ..., 6\}$ are the predictors and then T_2 is the target and $\{X_k, Y_k, U_k, k = 1, ..., 6\}$ are the predictors. In the second experiment the true model for T_2 includes interactions from the XOR function. In each experiment, 50 replicates of data sets are simulated, with $5120 = 10 * 2^9$ rows of data in each data set (so that all values of an attribute have the same

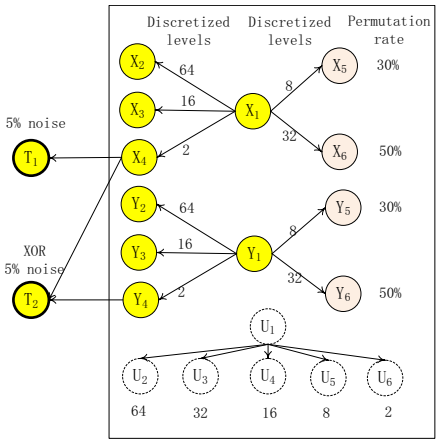


Fig. 1. Relationship between predictors and targets along with cardinalities. Here T_1 and T_2 denote the target attributes for the experiments with and without interactions, respectively.

number of rows). For example, for a two-value attribute, values 0 and 1 each have 2560 rows.

By designing such experiments, the order of the importance scores for the predictor attributes is known. In the first experiment: $Imp(X_1) = \ldots = Imp(X_4) > Imp(X_5) > Imp(X_6) > Imp(U_1) = \ldots = Imp(U_6)$. Therefore, there are four groups and attributes from the same group have equal information regarding T_1. In the second experiment: $Imp(X_1) = \ldots = Imp(X_4) = Imp(Y_1) = \ldots = Imp(Y_4) > Imp(X_5) = Imp(Y_5) > Imp(X_6) = Imp(Y_6) > Imp(U_1) = \ldots = Imp(U_6)$. Therefore, there are still four groups and attributes from the same group have equal information regarding T_2. A attribute importance measure should be able to indicate such orders of importance. We applied the original RF, chi-squared, OOBForest [10], cForest [9, 3] and pForest to each data set. Each forest used 200 trees. For the pForest test, we set $\delta = 20\%$ and $R = 200$. Because chi-squared works only for categorical attributes, the continuous predictors were removed before chi-squared importance measures were applied.

For the experiment without interactions Figure 2(a) illustrates the expected pattern. For basic RF in Figure 2(b), the importance measure prefers higher attributes for both informative and irrelevant attributes. Also, it can't discriminate between X_5 and X_6. Furthermore, the continuous attribute X_1 has the greatest importance score among the informative attributes. However, for the irrelevant attributes, the importance scores of the categorical attributes increase as the cardinality increases, and exceed the importance score of the X_1 when the cardinality equals 64. For cForest in Figure 2(c), there is no bias among the irrelevant attributes, which is consistent with the null case in [9]. cForest can also discriminate informative attributes from irrelevant attributes (though the differences between $Imp(X_5), Imp(X_6)$ and the U_k are not obvious), which is also consis-

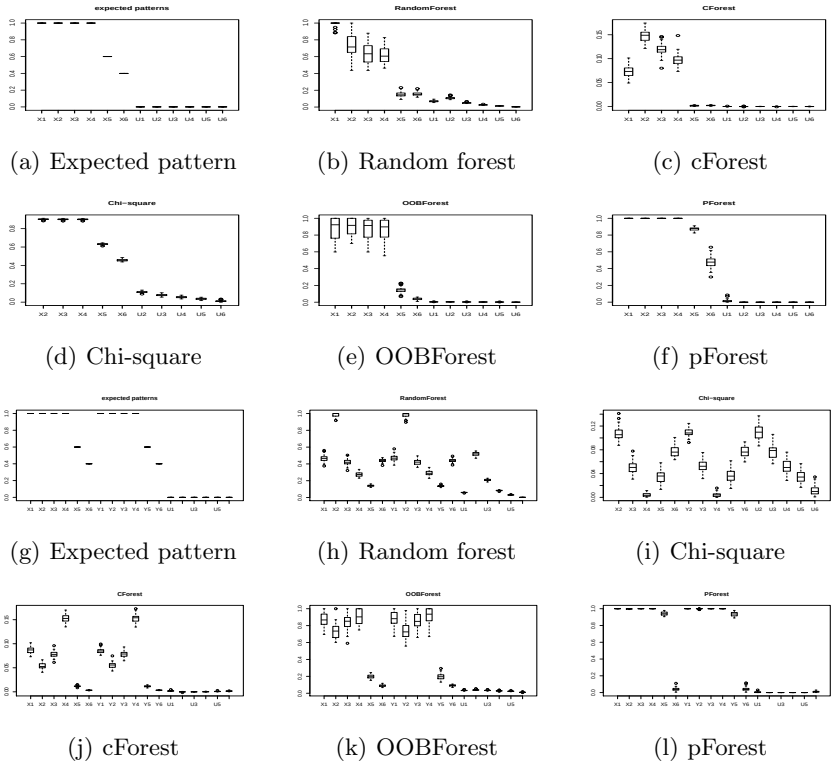
tent with the power case in [9]. However, for the informative attributes, cForest prefers higher cardinality attributes. Furthermore, it can not discriminate X_5 from X_6. For chi-squared in Figure 2(d), a concern is that higher-cardinality attributes are preferred for irrelevant attributes. For this experiment, it is able to rank informative attributes higher than irrelevant attributes and there is no obvious multi-valued problems for informative attributes. For both OOBForest in Figure 2(e) and pForest in Figure 2(f) there is no bias in both informative and irrelevant attributes. The expected orders among all predictor attributes are well preserved. Therefore, OOBForest has good performance here.

For the experiment with interactions Figure 2(g) illustrates the expected pattern. For RF in Figure 2(h)), the bias is even more severe than in the first experiment. RF cannot even discriminate irrelevant attributes from some informative attributes. The importance of U_2 is only less than X_2 and Y_2. Therefore, the attribute importance scores from RF are extremely unreliable here. Chi-squared in Figure 2(i) cannot even distinguish between the informative attributes and the irrelevant attributes. This is expected because chi-squared does not consider the interactions and this observation can clearly be extended to other methods (such as information gain), which only consider dependency between a single predictor and the target. For cForest in Figure 2(j) there is no obvious bias among the irrelevant attributes and cForest can also discriminate the informative attributes from the irrelevant attributes (although the importance difference between X_6 and $U_i s$ is not obvious). However, there is multi-valued problem in the informative attributes. In contrast to the previous experiment, cForest now prefers lower cardinality attributes. For OOBForest in Figure 2(k), there is no bias among the irrelevant attributes. The four groups can be discriminated. There are some minor importance differences among the most informative attributes. For pForest in Figure 2(l), there is no bias in both informative and irrelevant attributes. The expected orders among all predictors are well preserved.

From the experiments, RF is not reliable when predictors have different cardinality (prefers high cardinality attributes). cForest performs well for those irrelevant attributes with different cardinality. However, it is not reliable enough for the importance of informative attributes with different cardinalities. Without interactions, chi-squared prefers higher-cardinality ones for irrelevant attributes. More importantly, chi-squared is not reliable when interactions are present. The results of pForest are much better than RF. OOBForest also performs well in both experiments.

5 Conclusion

The bias of attribute importance measures is an important problem, and the common use of RF for attribute importance is shown to be a concern. We propose one method based on out-of-bag samples, while a second method uses the new concept of a partial permutation test to refine the attribute importance scores. The second method can be easily adapted to other feature scoring algorithms. We use a simulation framework that integrates different cardinalities, and where



(a) Expected pattern (b) Random forest (c) cForest

(d) Chi-square (e) OOBForest (f) pForest

(g) Expected pattern (h) Random forest (i) Chi-square

(j) cForest (k) OOBForest (l) pForest

Fig. 2. Feature importance for experiments without (Figures 2(a) to 2(f)) and with (Figures 2(g) to 2(l)) interactions. The X axis represents attributes, and the Y axis provides importance scores. Figures 2(a) and 2(g) illustrate the expected pattern for the no interaction and interaction cases, respectively (only relative expected measures).

non-equally informative attributes (with or without interactions) and irrelevant attributes co-exist. Our proposed methods are compared directly to two existing solutions for multi-value bias: chi-squared and a conditional inference framework and we observe that the existing methods are not always reliable for multi-valued predictors, while the proposed methods compare favorably in our experiments.

References

1. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics 26(10), 1340–1347 (May 2010)
2. Breiman, L., Friedman, J., Olshen, R., Stone., C.: Classification and Regression Trees. Wadsworth, Belmont, MA (1984)
3. Hothorn, T., Hornik, K., Achim, Z.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15, 651–674 (JAN 2006)

4. Igor, K.: On biases in estimating multi-valued attributes. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montral, Canada. pp. 1034–1040 (1995)
5. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom. pp. 249 – 256 (1992)
6. L, B.: Random forests. Machine Learning 45, 5–32 (Jan 2001)
7. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
8. Radivojac, P., Obradovic, Z., Dunker, A.K., Vucetic, S.: Feature selection filters based on the permutation test. In: Machine Learning: ECML 2004, 15th European Conference on Machine Learning (2004)
9. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(25) (JAN 2007)
10. Tuv, E., Borisov, A., Runger, G., Torkkola, K.: Feature selection with ensembles, artificial variables, and redundancy elimination. Journal of Machine Learning Research 10, 1341–1366 (2009)
11. White, A.P., Liu, W.Z.: Technical note: Bias in information-based measures in decision tree induction. Machine Learning 15(3), 321–329 (June 1994)