

Information-Theoretic Feature Selection for a Neural Behavioral Model

Bjorn Chambless, David Scarborough
Unicru, Inc.,
9300 SW Nimbus Ave.
Beaverton, OR 97008, USA
bchambless@unicru.com, dscarborough@unicru.com

Abstract

Employers of hourly workers typically experience high employee turnover. Due to costs associated with: training, hiring and termination, the overhead from with this high turnover rate is substantial. It is therefore desirable to construct employee selection procedures and analytic models to estimate the likely tenure of applicants for employment prior to a hiring decision. A critical component in the success of this effort to create a neural network model to estimate tenure was the application of information-theoretic feature selection. The benefits of this technique are demonstrated by comparison with results obtained using no feature selection and alternate methods of feature selection.

1 Introduction

Employee selection research attempts to link measures of individuals' characteristics to measures of job effectiveness for specific occupations. Industrial psychologists begin by analyzing a job and forming hypotheses about which characteristics increase the likelihood that an individual will be successful at a given type of work.

Measuring human characteristics and predicting human behavior is complex. Statistical prediction of work-related behavior requires expertise in differentiating individuals and a theoretical understanding of how such characteristics relate to behavior in the work place.

Despite nearly a century of applied research and numerous advances in measurement, modeling and research methods; behavioral prediction correlation coefficients do not approach the magnitude of those observed in the hard sciences. In a recent summary of hundreds of criterion validity studies in a wide variety of occupational settings[5], mean correlations between predictions based on biographical questionnaires and job performance was $\rho = 0.31$. Correlations between measures of personality traits and job performance tend to be even lower[8].

Due to the complexity of the problem domain, neural

networks would seem an excellent platform for behavioral models. Yet the vast majority of neural network research has been outside the behavioral sciences[7] and behavioral prediction models have continued to be developed "top-down" from cognitive theory, rather than "bottom-up" from empirical data. To treat behavioral prediction problems in the context of "data-mining" and "knowledge discovery" rather than testing the validity of *predetermined* theoretical models, has been uncommon. Early data mining efforts in industrial psychology were criticized (despite providing superior predictions) because a prediction came with, "...little understanding of underlying relationships or constructs accounting for the prediction." [11] These efforts were dismissed with the pejorative descriptions, "dust bowl empiricism" or "blind empiricism".

More recently, empirical exploration methods, including neural networks, have begun to gain acceptance as tools with which to discern underlying relationships. The point which eluded the critics of "dust bowl empiricism", was that an empirically established relationship creates incentive to generate a model to explain the phenomena. Somers[10] observed that neural networks, and other knowledge discovery methods are likely to uncover unanticipated relationship that have implications for theory development. Studying employee turnover, he used *self-organizing maps* to uncover a previously unobserved relationship between *job involvement*, *job performance* and *role conflict*. The correlation between job involvement and job performance is very low, yet in the relationship to role conflict there is a strong relationship between job involvement and job performance at *moderate* levels of job performance that *weakens* as role conflict increases. This unexpected finding illuminated a number of previously unexplained interaction effects observed, but not explained, by other researchers.

This is the first in a series of neural predictive models to provide employers of hourly workers information to in-

crease the efficiency and reliability of the hiring process. Future neural models will seek to predict behaviors such as: *sales*, *job abandonment* and *theft*.

This is not the first work demonstrating the use of information-theoretic methods for feature selection in conjunction with neural networks[1], nor is this the first application of neural model to the study of employee turnover[10] However, this may be the first work to integrate information-theoretic feature selection and neural modeling for behavioral prediction. Additionally, the feature selection method presented addresses certain potential shortcomings in algorithms suggested in previous works (see Section 2.2.3).

2 The Feature Selection Process

The source for the data used to develop this model is a large national video rental company. The sample contains 2048 cases, consisting of 160 responses to application questions collected prior to hiring and tenure (in days) for that former employee. The model was to generalize to the set of all potential applicants and accurately predict the length of employment for a given applicant, if hired. The application itself consists of 77 *bio-data* questions (these ask general, work related, information: job history, education and referrals) and 83 *psychometric* questions¹. The psychometric assessment portion was designed to predict the *reliability* of a applicant in an hourly, customer service position. For the purposes of model development, each question response was treated as a single feature and the reliability score was not provided to the neural network or feature selection algorithm.

2.1 Background

While any information gathered during the application process may have predictive value, it is best to reduce the set of input variables (independent variables or IVs). The justifications are the following

1. Not all potential IVs have significant predictive value. The use of variables with little or no predictive value as inputs adds *noise*. I.e. adding IVs to the model which cannot improve predictive capability may degrade prediction since the network will need to adapt to *filter* these inputs. This requires additional training time and neural resources.
2. Predictive models provide a mapping from an input space to an output space. The dimensionality of this input space increases with the number of inputs. Thus, the more parameters required to cover the mapping which in turn increases the *variance* of the model (in terms of the *bias/variance*

¹These are from the Unicru Customer Service Assessment test written by Dr. George Paajanen.

dilemma)[4]; a problem commonly referred to as the *curse of dimensionality*[2].

It is desirable to eliminate those IVs with less predictive power in favor of a less complex neural network model by applying *feature selection*. Such methods fall into two general categories [6]: *filters* and *wrappers*.

1. **Wrappers** use the relationship between model performance and IVs directly by iteratively experimenting with IV subsets. Since the nature of the bias of the feature selection method matches that of the modeling technique, this approach is theoretically optimal if the search is exhaustive.

Unfortunately, the exhaustive application of wrappers is computationally intractable for all but the smallest modeling problems since the number of possible subsets is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (1)$$

where n is the total number of IVs and k is the cardinality of the subset of features.

Additionally, there is the problem of non-determinism within the modeling process. In neural modeling, though training algorithms are typically deterministic, random initialization of the weight parameters varies the results of models developed with the same inputs. Therefore, even exhaustive trials may not prove conclusive with respect to estimating the predictive value of a set of features.

2. **Filters** analyze the relationship between sets of IVs and dependent variables (DVs) using methods independent of those used to develop the model.

One danger of this approach is that bias of the filter may be incompatible with that of the modeling technique. For example, a filter may fail to detect certain classes of constraint which the subsequent modeling stage may utilize. Conversely, the filter may identify relations which cannot be successfully modeled. Ideally, a filter would be completely inclusive in that no constraint which *might* be replicated by the subsequent modeling stage would be discarded.

2.2 An Information-Theoretic Approach to Feature Selection

Information-theoretic feature selection is derived from the statistical theory of *independent events*. Events p_1, p_2, \dots, p_n are considered *statistically independent* if and only if the probability P , that they all occur on a given trial is

$$P = \prod_{i=1}^n p_i$$

The degree to which a joint distribution of probabilities diverges from the independence distribution may be used as a measure of the statistical *dependence* of the events.

Information-theoretic *entropy*²[9] provides a convenient metric for quantifying the difference between distributions. The entropy, $H(X)$ (measured in bits), of the distribution of a discrete random variable, X , with n states is

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$$

where p_i is the probability of state i .

Entropy is maximized when a distribution is most uncertain. If a distribution is discrete, this occurs when it is uniform. Figure 1 shows a graph of the entropies of a single variable, 2-state distribution as the state probabilities vary.

For a multivariate distribution constrained by fixed marginals, the distribution which *maximizes* entropy will be the independence distribution (calculated as the product of the marginals). The distribution which *minimizes* entropy is the distribution for which the variables are completely dependent.

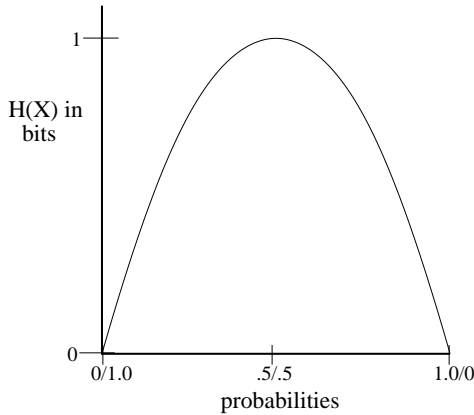


Figure 1: Entropy for X , a 2-state discrete distribution

Dependence is *constraint* between variables, so as constraint is reduced, entropy increases. Information-theoretic analysis can therefore be used to measure constraint. (see Figure 2) For a joint distribution of discrete variables, X and Y , the total entropy, $H(XY)$ is

$$H(XY) = - \sum_{i,j} p_{ij} \log_2 p_{ij}$$

²Information-theoretic entropy should not to be confused with its thermodynamic counterpart.

where p_{ij} is the probability of state i, j occurring in the joint distribution of X and Y , where i designates the state of X and j is the state of Y . The entropies of X and Y are computed with the marginals of the joint distribution

$$H(X) = - \sum_i \left(\sum_j p_{ij} \right) \log_2 \left(\sum_j p_{ij} \right)$$

$$H(Y) = - \sum_j \left(\sum_i p_{ij} \right) \log_2 \left(\sum_i p_{ij} \right)$$

Information transmission (also known as *mutual information*.) is the measure of the distance between the independence and observed distributions along the continuum shown in Figure 2. For X and Y , $T(X:Y)$ (the information transmission between X and Y), is computed

$$T(X:Y) = H(X) + H(Y) - H(XY) \quad (2)$$

In a directed system, the measure of information transmission between the distribution of an independent variable X and a dependent variable Y is a gauge of the predictive value of X . $H(X) + H(Y) = H(XY)$ if and only if there is no constraint between X and Y , in which case X would be a poor predictor for Y .

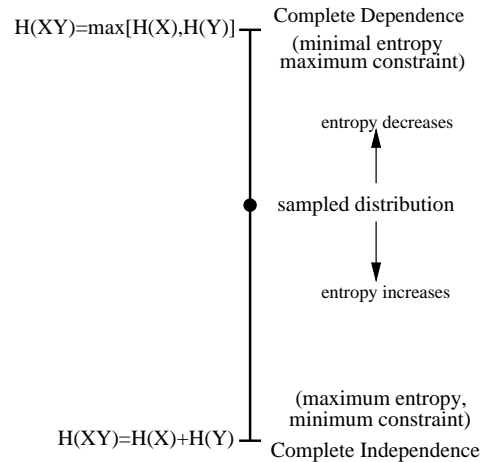


Figure 2: Constraint vs. entropy

2.2.1 Statistical Significance of Transmission Calculations. In order for a computed transmission value, T , to be considered an accurate measure of existing constraint, the statistical significance of T for some confidence level, α , must be determined using the well-known χ^2 test. The degrees of freedom (df) for a transmission, $T(X:Y)$, is calculated

$$df_{T(X:Y)} = df_{XY} - df_X - df_Y \quad (3)$$

As the size of the joint distribution increases, so does the df for the significance of the transmission value. Since χ^2 significance decreases as df increases, the data requirements for transmissions containing a large number of variables quickly becomes overwhelming.

2.2.2 Determining the Optimal Feature Set.

The goal is to discover a subset S of the independent variables V which has the same predictive power as the entire set with respect to the dependent variables, D .

$$T(V:D) \approx T(S:D)$$

The filtering process should therefore be guided by the following³

1. if S' is any subset of V smaller than S , then $T(S':D)$ is significantly smaller than $T(S:D)$.
2. if S' is any subset of V larger than S , then $T(S':D)$ is not significantly larger than $T(S:D)$.

2.2.3 Higher-Order Interactions. Higher-order interactions are *synergies* between variables where the predictive power of a set of variables is significantly higher than that of the sum of the individual variables. In terms of information transmission for the IVs X_1, \dots, X_n , and dependent variable D , this is represented,

$$T(X_1:D) + \dots + T(X_n:D) < T(X_1, \dots, X_n:D)$$

An illustration of this phenomenon among discrete binary variables: A, B and C , is shown by the contingency table⁴ in Figure 3.

		C=0	
A	0	1/4	0
	1	0	1/4
		0	1
		B	

		C=1	
A	0	0	1/4
	1	1/4	0
		0	1
		B	

Figure 3: Contingency table for distribution ABC

For this system, the following transmissions are computed

$$\begin{aligned} T(A:C) &= H(A) + H(C) - H(AC) = 0 \text{ bits} \\ T(B:C) &= H(B) + H(C) - H(BC) = 0 \text{ bits} \\ T(AB:C) &= H(AB) + H(C) - H(ABC) = 1 \text{ bit} \end{aligned}$$

Knowledge of A or B individually does nothing to reduce the uncertainty of C , but knowledge of A and B

³These guidelines are similar to criterion in Extended Dependency Analysis (EDA). See [3]

⁴Frequencies have been replaced by their percentage of the joint distribution

eliminates *all* uncertainty since only one state of C is possible. With only first order transmissions values, A and B would not appear to be predictive features, when in fact, together they are ideal.

Higher order interactions were observed in the video clerk tenure data. Table 2.2.3 lists the top ten single variable transmissions between the psychometric questions and tenure. Table 2.2.3 shows the top five, two and three variable transmissions. Each of the most predictive sets of questions (based on transmission values) in both the second and third order lists, $T(q35\ q73:tenure)$ and $T(q4\ q12\ q39:tenure)$, contain only one question from the top ten most predictive questions based on first order transmissions.

variables	trans.	%H(DV)	df	χ^2 sig.
T(q83:tenure)	0.0168	0.754	27	0.999
T(q3:tenure)	0.0140	0.628	27	0.991
T(q63:tenure)	0.0135	0.607	27	0.987
T(q65:tenure)	0.0133	0.598	27	0.985
T(q48:tenure)	0.0133	0.595	27	0.984
T(q44:tenure)	0.0132	0.593	27	0.984
T(q35:tenure)	0.0128	0.573	27	0.977
T(q21:tenure)	0.0127	0.569	27	0.975
T(q8:tenure)	0.0123	0.553	27	0.967
T(q69:tenure)	0.0123	0.552	27	0.966

Table 1: Single order transmissions between psychometrics and tenure

variables	trans.	%H(DV)	df	χ^2 sig.
T(q35 q73:tenure)	0.0593	2.663	135	1.00
T(q21 q83:tenure)	0.0588	2.639	135	1.00
T(q39 q65:tenure)	0.0585	2.627	135	1.00
T(q61 q70:tenure)	0.0569	2.553	135	0.999
T(q44 q53:tenure)	0.0567	2.546	135	0.999
T(q4 q12 q39:tenure)	0.1808	8.112	567	0.921
T(q10 q39 q65:tenure)	0.1753	7.864	567	0.811
T(q4 q39 q44:tenure)	0.1720	7.718	567	0.712
T(q4 q39 q51:tenure)	0.1718	7.709	567	0.705
T(q52 q61 q70:tenure)	0.1717	7.702	567	0.700

Table 2: Higher (second and third) order transmissions between psychometrics and tenure

Such interactions complicate the search for the optimal set S since the members of V may not appear as powerful predictors in calculated transmissions using sets of features of cardinality less than $|S|$ (the cardinality of the optimal subset S).

Unfortunately, due to issues of χ^2 significance, it is frequently impossible to calculate significant transmission values for sets of variables of cardinality approaching $|S|$. Additionally, since the number of subsets of a given cardinality soon become very large (see equation 1), even if the significance issues were addressed, computational limitations would persist.

In feature selection algorithms which approximate an exhaustive search for S by computing only pairwise

transmissions[1], higher-order interaction effects are not detected. Such methods may not accurately approximate S since only variables which are strong single variable predictors will be selected.

2.2.4 Heuristics for Information-Theoretic Feature Selection. Based on the following arguments, heuristics were applied in an effort to address the problems of combinatorics and significance in measuring higher-order relations.

It is hypothesized that, although it is possible for members of the optimal subset of IVs, S , to be completely absent from all large lower order transmissions, this is unlikely. An omission should be increasingly unlikely as the order of the transmissions calculated approaches $|S|$. It is therefore assumed that all members of S will appear in the top n transmissions of the highest order transmission computed, where n is sufficiently large. I.e. it is assumed that as $n \rightarrow |S|$, the *union* of the set of IVs appearing in the most predictive transmissions will approach S .

With these assumptions, an algorithm for generating an approximation to S (S') given the set V of all IVs and the set D of all DVs, is presented.⁵

In the following algorithm, T_k will be used to denote the set of transmissions of order k (containing k IVs) from a set of n features.

1. Calculate the transmissions, T_k , for the highest order, k , for which all $\binom{n}{k}$ transmissions may be calculated.
2. Choose the m unique transmissions of the greatest magnitude from T_k to be the base set for higher-order transmissions.
3. Generate T'_{k+1} by adding the IV to each member of T_k which generates the set T_{k+1} with the largest transmission values. Note that T'_{k+1} is a *subset* of T_{k+1} since it contains only those members of T_{k+1} which can be generated from T_k by adding one independent variable to each transmission.
4. Discard any duplicate transmissions.
5. Repeat Steps 3 and 4 until χ^2 significance is exhausted⁶.

⁵This algorithm assumes computational limits will be reached before limits of statistical significance.

⁶The common "rule of thumb" with respect to χ^2 significance is that, on average, the population of a given cell should be 5. Therefore the largest joint distribution for which transmission can be calculated should have no more than $D/5$ cells where D is the size of the data sample

6. Take the union of the variables appearing in as many of the most predictive transmissions as is necessary to generate a set of size $|S|$. This union is S' , the approximation of the set S .

Since $|S|$ is unknown, this value must be estimated. However, $0 \leq |S| \leq |V|$, so it is often feasible to experiment with the S' for each cardinality.

2.2.5 Dependence Between Features. An issue raised by other feature selection algorithms[1] is the effect of dependence *between* members of S' . This dependence may be viewed as the *redundancy* in the predictive content of the variables. One solution proposed is to calculate all pairwise transmissions, $T(s'_i : s'_j)$, between features s'_i and s'_j from a candidate S' . Features which exhibit high dependence (high pairwise transmission) are penalized with respect to the likelihood of their inclusion in the final S' .

Dependence between features is dealt with *implicitly* in the algorithm presented in Section 2.2.4 since such dependence will reduce the entropy, thereby reducing the magnitude of the transmission between a set of features and the set of dependent variables (see Equation 2). Highly redundant feature sets will have low transmission values relative to less redundant sets of the same cardinality and will therefore be less likely to contribute to S' .

3 Data Clustering

While *tenure in days* is a discrete measure, the number of possible states makes it difficult to use the variable without transformation since a large number of states makes the joint distribution sparse (high df relative to the data population. See Equation 2.2.1.) and any transmissions calculated statistically insignificant. Since *tenure* is an ordered variable, applying a clustering algorithm was not problematic.

Clustering is a form of *compression*, so care must be taken to minimize information loss. The clustering phase was guided by efforts to maximize the entropy of the clustered variable within the confines of the needs of statistical significance.

Though transmission values did vary across clustering algorithms and granularity, the results in terms of S' were consistent. Unfortunately, a full account of the clustering procedures employed is beyond the scope of this paper.

4 Experimental Results

Transmissions were calculated using software developed by the authors which combined cluster analysis and information-theoretic analysis. For the video clerk data

set (containing 160 IVs), it was decided that the cardinality of the sets of IVs for which all transmissions could be calculated was 4. From there, two additional orders of cardinality were calculated by supplementing the 4th order transmissions (as described in step 3 of the algorithm). The union of independent variables appearing in the largest transmissions was taken to be S' . Experimentation with neural models using S' of different cardinalities yielded the best results when $|S'| = 56$.

An interesting aspect of the application questions chosen by the feature selection method was the mix of bio-data and psychometrics. Of the 56 features used as inputs to the most successful model, 31 came from the bio-data section of the application and 25 came from the psychological assessment. Of particular interest was the “coupling” of certain bio-data and assessment questions. Such pairs would appear together throughout the analysis of transmission over a range of cardinalities. I.e. they would appear as a highly predictive pair and would subsequently always appear together in higher-order sets of IVs.

The synergistic effect between the two classes of question became apparent when models were generated using exclusively one class or the other (using *only* psychometrics or *only* bio-data questions). With comparable numbers of inputs, these models performed significantly *worse* than their more diverse counterparts. These results are particularly interesting since psychological assessments typically do *not* include responses from such diverse classes of questions.

4.1 Model Performance

The most successful neural model developed was a single hidden layer, feed-forward neural network with 56 inputs ($|S'| = 56$), and 40 hidden nodes. The network was trained using the *conjugate gradient* method. Of the total data set size of 2084, 1784 were allocated to the training set and 300 were “hold-out”.

As mentioned in Section 1, the performance measures of behavioral prediction models are typically measured using the correlation coefficient. For the neural model described, the correlation between prediction and actual tenure for the hold-out sample was $\rho = 0.51$. For comparison, a number of other models were generated using either no feature selection or alternate feature selection methods. These models used the same network architecture and training algorithm. The best model generated using the entire data set (all features), was a 160-90-1 configuration (160 inputs and 90 hidden layer nodes) which achieved a maximum hold-out correlation of $\rho = 0.44$. Alternate feature selection algorithms: *ge-*

netic algorithms, and *forward* and *reverse* stepwise regression, using the same number of features (56), failed to achieve a hold-out correlation better than $\rho = 0.47$.

5 Conclusion

The findings presented suggest that information-theoretic feature selection is a viable and accurate method of identifying predictors of job performance in employee selection. The capacity to identify non-linear and higher-order interactions ignored by other feature selection methods represents a significant step forward in the development of neural network based predictive behavioral models.

References

- [1] Battiti, R., “Using mutual information for selecting features in supervised neural net learning”, *Neural Networks*, 5, 537-550, 1994
- [2] Bellman, R., *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961
- [3] Conant, R., “Extended Dependency Analysis of Large Systems”, *Int. J. General Systems*, vol 14, pp. 97-141, 1988
- [4] Geman, S., E. Bienenstock, and R. Doursat, “Neural Networks and the Bias/Variance Dilemma”, *Neural Computation*, MIT Press, Vol. 4, pp 1-58, 1992
- [5] Hough, L.M., F.L. Oswald, “Personnel Selection: Looking Toward the Future-Remembering the Past”, *Annual Review of Psychology*, vol 51, pp. 631-664, 2000
- [6] John, G.H., R. Kohavi, K. Pfleger, “Irrelevant Features and the Subset Selection Problem”, *Proceedings of the 11th International Conference on Machine Learning*, 1994
- [7] Rappa, M.A., K. Debackere, “Technological Communities and the Diffusions of Knowledge”, *R & D Management*, 22(3), pp. 209-220, 1992
- [8] Schmitt, N., R.Z. Gooding, R.A. Noe, M. Kirsh, “Meta-analysis of Validity Studies Published between 1964 and 1982 and investigations of study characteristics”, *Personnel Psychology*, 37-3, pp. 407-422, 1984
- [9] Shannon, C.E., “A mathematical theory of communication.” *Bell Sys. Tech. Journal*, vol 27, 1948
- [10] Somers, M.J., “Application of Two Neural Network Paradigms to the Study of Voluntary Employee Turnover”, *Journal of Applied Psychology*, vol 84, 1999
- [11] Stokes, G.S., M. Mumford, W. Owens editors, *Bio-data Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction*, CPP books, Palo Alto, CA, 1994