TF-IDF Weighted Similarity Estimates for Unseen Categories

David Bang
Ancestry
153 Townsend Avenue
San Francisco, CA
dbang@ancestry.com

Feng-Chang Lin, Michael R. Kosorok
Univ. of North Carolina at Chapel Hill
135 Dauer Dr.
Chapel Hill, NC
flin33@unc.edu

Alexander Comerford
Bloomberg
731 Lexington Ave
New York, NY
acomerford3@bloomberg.net

Abstract

Results of machine learning models for clinical data are difficult to generalize outside of the context from which the data was gathered due to substantial differences in the target population with respect to geography, demography, etc. Naturally applying a prediction model in a new context becomes problematic since the train and test data differ in distribution and a common pitfall in such a procedure is handling unseen categories. This is exacerbated in electronic health records data due to the large set of possible categorical occurrences such as ICD-10 and CPT codes. Modern approaches rely on preprocessing techniques such as imputation by treating unseen categories as missing values or by assigning them predetermined clusters. TF-IDF Similarity Weighted Estimates (TIWS) is a novel framework by treating categorical data in an NLP context. TIWS assigns the unseen category a linear combination of seen categories with weights based on similarity measures.

Introduction

Generalizability is an issue in predictive modeling. It is standard practice to have the train and test sets to have the same distribution so that models can be robust in their estimation. However, there are several instances in which unseen data deviates from the train data due to a new category or a distribution shift for a feature, etc. Current methods address this issue as an imputation problem or simply assigning the unseen category a predetermined cluster (Jin X. 2011). However, imputation methods specify a single observation rather than an entire category and it is tedious work developing clusters for preprocessing. Our method addresses this issue by developing a novel technique for deriving estimates for unseen categories which allows for both demonstrable improvements in prediction and causal inference. The method involves subsetting the data into its categorical partitions and then using information across and between each category through a customized TF-IDF encoding as similarity weights for the target statistic estimate of the unseen category.

Results were compared against common imputation techniques such as kNN and Bayesian ridge across case studies. In preliminary results, the method performed better than kNN which is an improvement due to computational efficiency, aggregated inference, and nonparametric modeling. Data used to derive preliminary results include the Titanic and Wake County Sudden Death data.

Method

Suppose we have train and test data $\mathcal{D} = \{(x, y)\}_{i=1}^{n}$, $\mathcal{D}^* = \{(x^*, y^*)\}_{i=1}^{n^*}$, respectively. Let $x_i = (x_{i1}, \ldots, x_{ip})^T \in R^p$ consists of only categorical features where x_{ij} represents the j^{th} feature of x_i, and $j = 1, ..., p$. Let c_{jk} represent the k^{th} category of the j^{th} feature, and let $k = 1, .., l_j$ with l_j denoting the number of known categories in the j^{th} feature. WLOG, we assume there is only one unseen category in the feature j. We let $c_{jl_{j+1}}$ represent the unseen category. We compute a target estimate $\hat{y}_{jk}$ (Dorogush et al. 2018) for the target parameter $E(y|x_{ij} = c_{jk})$ using the training label y. An empirical average of y with the same category c_{jk} can be used.

$$\hat{y}_{jk} = \frac{\sum_{i=1}^{n} I(x_{ij} = c_{jk}) \cdot y_i + ap}{\sum_{i=1}^{n} I(x_{ij} = c_{jk}) + a}$$

where $a > 0$ is a parameter. A common setting for p is the average target value in the dataset. This assigns categories a numeric value. To construct documents of categories it is important to combine only the feature data $\mathcal{D}_x \bigcup \mathcal{D^*}_x = \{(\tilde{x},)\}_{i=1}^{n+n^*}$. A document of a specific category is defined by

$$D_{jj'} = \begin{array}{c} \\ c_{j1} \\ c_{j2} \\ \vdots \\ c_{jl_{j+1}} \end{array} \begin{array}{c} \begin{array}{ccc} c_{j'1} & \cdots & c_{j'l_{j'}} \end{array} \\ \begin{pmatrix} f(c_{j1}, c_{j'1}) & \cdots & f(c_{j1}, c_{j'l_{j'}}) \\ f(c_{j2}, c_{j'1}) & \cdots & f(c_{j2}, c_{j'l_{j'}}) \\ \vdots & \vdots & \vdots \\ f(c_{jl_{j+1}}, c_{j'1}) & \cdots & f(c_{jl_{j+1}}, c_{j'l_{j'}}) \end{pmatrix} \end{array}$$

where $f(c_{jk}, c_{j'k'}) = \sum_{i=1}^{n} I(\tilde{x}_{ij} = c_{jk}) I(\tilde{x}_{ij'} = c_{j'k'})$ for $j' \neq j$, $j' = 1, ..., p$, $k' = 1, ..., l_{j'}$, and $k = 1, ..., l_j$. In short, frequency counts where $\tilde{x}_{ij} = c_{jk}$ and $\tilde{x}_{ij'} = c_{j'k'}$. We consider $\mathcal{D}_{jj'}$ as a text corpus for the j^{th} category, where its words are represented by each c_{jk}. For $j' = 1, \ldots, p$

and $j' \neq j$, we augment the $D_{jj'}$ matrix to the full document matrix D_j that combines matrices by categories of feature j. The document matrix D_j can be defined by $D_j = (D_{j1}|\cdots|D_{jp})$ where D_j is a $l_{j+1} \times \sum_{j'=1, j' \neq j}^{p} l_{j'}$ matrix. Now we create the term frequency (TF) and inverse document frequency (IDF) matrices as follows:

$$T_{D_j}(c_{jk}, c_{j'k'}) = \frac{f(c_{jk}, c_{j'k'}) - \mu_{c_{jk}}}{\sigma_{c_{jk}}}$$

$$I_{D_j}(c_{jk}, c_{j'k'}) = T_{D_j^T}(c_{j'k'}, c_{jk})^T$$

where $\mu_{c_{jk}} = \frac{1}{n_j} \sum_{\substack{j'=1 \\ j' \neq j}}^{p} \sum_{k'=1}^{l_{j'}} f(c_{jk}, c_{j'k'})$,

$\sigma^2_{c_{jk}} = \frac{1}{n_j - 1} \sum_{\substack{j'=1 \\ j' \neq j}}^{p} \sum_{k'=1}^{l_{j'}} \{f(c_{jk}, c_{j'k'}) - \mu_{c_{jk}}\}^2$,

$n_j = \sum_{\substack{j'=1 \\ j' \neq j}}^{p} l_{j'}$. TF is the distribution of words that describe the document and IDF is the impact factor of those words. IDF penalizes useless words.

Our proposed TIWS is the common standardization procedure for both TF and IDF due to its ubiquitious use but they can be uniquely defined depending on the context. Let $g(A, B)$ be a transformation function that aggregates two matrices A and B. Let $H_{D_j} = g(T_{D_j}, I_{D_j})$ be the aggregation matrix which uses the Hadamard (element-wise) multiplication. One can utilize a similarity metric $s(c_{jk}, c_{jk'})$ to describe the similarity between two row vectors of H_{D_j} at categories c_{jk} and $c_{jk'}$, and create a symmetric similarity matrix S that includes all pairwise similarities. Here, we choose a modified cosine similarity(B. and L. 2013), which is defined as

$$s(c_{jk}, c_{jk'}) = \frac{1}{2} + \frac{1}{2} \cdot \frac{H_{D_j}(c_{jk}) H_{D_j}(c_{jk'})^T}{\sqrt{H_{D_j}(c_{jk})^{\otimes 2}} \sqrt{H_{D_j}(c_{jk'})^{\otimes 2}}}$$

where $H_{D_j}(c_{jk})$ is the row vector of H_{D_j} matrix at category c_{jk}, and $a^{\otimes 2} = aa'$ for a row vector a. It is recommended that $s(c_{jk}, c_{jk'}) \in [0, 1]$ since the similarity coefficients will be used as weights to find the predicted value of the target parameter of the unseen category. Finally, the unseen category $c_{jl_{j+1}}$ is defined by

$$\hat{y}_{jl_j+1} = \frac{\sum_{k=1}^{l_j} s(c_{jk}, c_{jl_{j+1}}) \cdot \hat{y}_{jk}}{\sum_{k=1}^{l_j} s(c_{jk}, c_{jl_{j+1}})}$$

Multiple Unseen Categories

Suppose there are multiple unseen categories in the test set for the jth feature. Then we only need to be working with the document matrix D_j and its similarity matrix S_j. When deriving the D_j it is important to note that we are taking frequency counts of all c_{jk} for $j = 1, ..., p$ and $k = 1, ..., l_j$ in $\mathcal{X}$. However, a TIWS estimate for an unseen $\hat{y}_{jl_{j+a}}$ for does not use any information from any other unseen category $c_{jl_{j+b}}$ for $a = 1, ..., m$ and $b = 1, ..., m$ where $a \neq b$ and for $m = 1, .., \infty$ unseen categories. The derivation is described in the following:

We set $diag(S_j)$=0 or $s(c_{ji}, c_{jk})$=0 for $i = k$ for $i = 1, ..., l_{jl_{j+m}}$ and $k = 1, ..., l_{jl_{j+m}}$ where $dim(S_j) = l_{j+m-1}$ x l_{j+m-1} and $\vec{y} = [\hat{y}_{j1}, \ldots \hat{y}_{jl_j}, 0, \ldots 0]'$. We estimate the vector of unseen categories for the jth feature as follows:

$$\vec{c} = S_j \cdot \vec{y}$$

$$\vec{c} = \begin{bmatrix} \sum_{i \neq 1, i=1}^{l_{j+m}} s(c_{j1}, c_{ji}) \hat{y}_{ji} \\ \sum_{i \neq 2, i=1}^{l_{j+m}} s(c_{j2}, c_{ji}) \hat{y}_{ji} \\ \vdots \\ \sum_{i \neq l_{j+1}, i=1}^{l_{j+1}} s(c_{jl_{j+1}}, c_{ji}) \hat{y}_{ji} \\ \sum_{i \neq l_{j+m}, i=1}^{l_{j+m}} s(c_{jl_{j+m}}, c_{ji}) \hat{y}_{ji} \end{bmatrix}$$

Notice that $\vec{c}_{[l_{j+1}]} = \sum_{i=1}^{l_{j+m}} s(c_{jl_{j+1}}, c_{ji}) \hat{y}_{ji}$

$= \sum_{\substack{i \neq l_{j+1} \\ i=1}}^{l_{j+m}} s(c_{jl_{j+1}}, c_{ji}) \hat{y}_{ji}$ since we set $s(c_{ji'}, c_{ji})$=0 for $i = i'$ for $i = 1, ..., p'$ and for $i' = 1, ..., p$. Also notice that $\vec{c}_{[l_{j+1}]}$ does not contain $\hat{y}_{jl_{j+1}}$. To get the TIWS estimate for $\hat{c}_{jl_{j+1}}$ simply equate $\hat{y}_{jl_{j+1}} = \vec{c}_{[l_{j+1}]}$. So for multiple unseen categories $\{\hat{y}_{jl_{j+1}}, ..., \hat{y}_{jl_{j+m}}\}$ we can simply grab $\vec{c}_{[l_{j+1}:l_{j+m}]}$ for those TIWS estimates. Distribute the remaining weights that were set to 0 equally among all reference categories i.e. distribute 1 - $\frac{1}{m} \sum_{\substack{i \neq l_{j+1} \\ i=1}}^{l_{j+m}} s(c_{jl_{j+1}}, c_{ji}) \hat{y}_{ji}$. The case of multiple unseen categories across multiple features is simply an iterative extension.

Algorithmic Complexity & Relation to kNN

It can be shown that TIWS is identical to kNN (Fix and Hodges 1951) under certain conditions. Namely, $kNN_{c_{jl_{j+1}}} = \frac{\sum_{k=1}^{l_j} p_{jk} \cdot \hat{y}_{jk}}{\sum_{k=1}^{l_j} p_{jk}}$. Generally, kNN uses information within a local bound in an iterative fashion whereas TIWS uses information in a one-shot aggregate fashion. The time complexity for TIWS is $O((n + n*) \cdot l_{j+m})$ and space complexity of $O((n + n*) \cdot l_{j+m})$ for the temporary similarity matrix S.

Results

Sudden Death ICD-10

We benchmarked predictive performances for TIWS, kNN and Bayesian ridge imputations. ICD-10 codes are encoded diagnosis determined by a medical professional. The full ICD-10 code description for first, second, and third diagnostics were used. There are 89 unique ICD-10 five letter codes for the first diagnostic. This scenario is an edge case

since the majority of each categorical ICD-10 instance contains only one instance. Only 14 out of the 89 ICD-10 codes contain more than one observation. This is a special type of sparsity since one can imagine that the majority of data consists of outliers. Recall that TF-IDF attempts to capture the frequency of each category with respect to categories within the same feature and with categories across multiple features. The document matrix for this data largely consists of frequency count values of 1. It is interesting to see how the methods perform.

The results in Table 1 demonstrate that all of the imputation methods performed similarly. Although TIWS performed slightly worse in the XGBoost(Chen and Guestrin 2016) scenario, we can see that even with a lack of information, TIWS performs as solidly as the other imputation methods. This result gives a compelling reason to consider TIWS since it performed similarly to its peers under extreme circumstances. It would be interesting to see how future ICD-10 codes will cluster to the 89 first diagnostics ICD-10 codes compared to the general medical classification groups outlined by the World Health Organization (WHO).

We also considered the case of using the predetermined letter cluster to compare how the various methods performed. The first few characters of the ICD-10 code correspond to a general class of medical diagnosis outlined by WHO. We scaled back from working with 156 categories to 18 general categories. The results in Table 1 demonstrate that once again all of the imputation algorithms perform similarly with TIWS performing slightly better. It is interesting to note that the kNN and Bayesian Ridge algorithms all selected the same imputation decisions whereas TIWS slightly deviated from those choices.

Titanic Cluster Analysis

There is difficulty in assessing how well TIWS estimates reflect the ground truth in the ICD-10 codes case due to the sheer number of potential diagnoses. So the popular Titanic data set will be evaluated to understand how TIWS is determining its estimates across multiple cases. The cases considered include multiple unseen categories for a single feature, multiple unseen categories across multiple features, and diverging performances among imputation methods. Due to the spacing limit, we only demonstrate the case where TIWS diverges in its categorical selection from all the kNN models. Note that for all the other cases analyzed, TIWS outperformed kNN.

An interesting case occurred when all of the kNN models selected the same decision path but TIWS chose otherwise. The results are shown in Table 2. All observations that fell under the category $Q \in embarked$ were removed from the train set and used as the test set. It is important to note that $embarked$ consisted of only three categories $\{S, C, Q\}$. For every observation, $embarked_Q$ was most similar to $embarked_S$ for the kNNs, but TIWS determined that $embarked_Q$ was most similar to $embarked_C$. A reason for this may be the number of observations per class $\{S, C, Q\} \rightarrow \{644, 168, 77\}$. So there was a much higher chance that the majority of the nearest neighbors consisted of $embarked_S$. TIWS performed significantly better across all metrics and folds. This highlights a limitation of kNN. kNN can be very near-sighted since it does not consider a holistic analysis of the data. This can also be viewed as empirical evidence of our proposition that TIWS approximates kNN under certain conditions namely when the similarities are identical to that of the proportion of the available seen categories which occurs when neighbors = n.

Conclusion

Compared to imputation techniques, TIWS gains an advantage in its estimation through an aggregated estimate rather than through each individual observation. This helps answer questions regarding how an unseen category behaves in relation to a reference of seen categories. In addition, TIWS can be automatically used to estimate cases for when there are multiple unseen categories across any number of features. TIWS performed strongly under sparsity of information for numerous categories in the ICD-10 case and was able to outperform kNN by picking up global nuances among categories in the Titanic data. However, there are a few drawbacks in TIWS estimation. TIWS is static in nature for several algorithms i.e. an unseen category will follow the decision path of its most similar seen category for decision trees, boosting, etc. Also, proving asymptotic properties for TIWS statistics are difficult since we are considering it in an NLP context. Despite these drawbacks, one should consider using TIWS if there is expected to be a large number of new categories.

References

B., L., and L., H. 2013. Distance weighted cosine similarity measure for text classification. *Intelligent Data Engineering and Automated Learning – IDEAL 2013*.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. 785–794.

Dorogush, A.; Gulin, A.; Gusev, G.; Kazeev, N.; Prokhorenkova, L.; and Vorobev, A. 2018. Catboost: Unbiased boosting with categorical features. *In Advances in Neural Information Processing Systems*.

Fix, E., and Hodges, J. 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. *USAF School of Aviation Medicine, Randolph Field*.

Jin X., H. J. 2011. K-means clustering. *Encyclopedia of Machine Learning*.

ICD-10 Prediction Analysis							
Letter Codes (18 categories)				Full Codes (156 categories)			
Random Forest							
Method	Accuracy	Precision	Recall	Method	Accuracy	Precision	Recall
knn(20)	0.700	0.744	0.842	knn(20)	66.2%	73.0%	78.4%
knn(40)	0.700	0.744	0.842	knn(40)	66.9%	73.4%	78.9%
knn(80)	0.700	0.744	0.842	knn(80)	66.5%	73.1%	78.9%
Bayes	0.700	0.744	0.842	Bayes	67.2%	73.9%	78.9%
TIWS	0.704	0.747.	0.842	TIWS	67.2%	73.9%	78.9%
XGBoost							
Method	Accuracy	Precision	Recall	Method	Accuracy	Precision	Recall
knn(20)	0.718	0.752	0.857	knn(20)	65.8%	73.0%	76.8%
knn(40)	0.718	0.752	0.857	knn(40)	66.2%	73.1%	77.3%
knn(80)	0.718	0.752	0.857	knn(80)	66.5%	72.8%	77.3%
Bayes	0.718	0.752	0.857	Bayes	66.5%	73.3%	77.8%
TIWS	0.722	0.756	0.857	TIWS	65.1%	72.5%	76.3%

Table 1: ICD-10 Prediction Analysis for letter and full codes

Diverging performance for TIWS & kNN					
Unseen Categories $\{embarked_Q\}$ (n=77)					
CV1 (n=51)		CV2 (n=51)		CV3 (n=52)	
Method	Results	Method	Results	Method	Results
kNN (all)	Accuracy: 70.6% Precision: 60.0% Recall: 35.3%	kNN (all)	Accuracy: 68.6% Precision: 63.6% Recall: 36.8%	kNN (all)	Accuracy: 67.3% Precision: 81.8% Recall: 37.5%
TIWS	Accuracy: 80.4% Precision: 68.4% Recall: 76.5%	TIWS	Accuracy: 74.5% Precision: 63.6% Recall: 73.7%	TIWS	Accuracy: 78.8% Precision: 76.0% Recall: 79.2%

Table 2: General unseen categories for a specific feature 'embarked'