REMOVING DATA WITH NOISY RESPONSES IN REGRESSION ANALYSIS

Alan Wisler, Visar Berisha*, Karthikeyan Ramamurthy†, Andreas Spanias*‡, Julie Liss**

ABSTRACT

In regression analysis, outliers in the data can induce a bias in the learned function, resulting in larger errors. In this paper we derive an empirically estimable bound on the regression error based on a Euclidean minimum spanning tree generated from the data. Using this bound as motivation, we propose an iterative approach to remove data with noisy responses from the training set. We evaluate the performance of the algorithm on experiments with real-world pathological speech (speech from individuals with neurogenic disorders). Comparative results show that removing noisy examples during training using the proposed approach yields better predictive performance on out-of- sample data.

Index Terms— robust regression, outlier removal, noisy data, minimum spanning tree, Friedman-Rafsky statistic

1. INTRODUCTION

Instance pruning is a long-standing problem in the field of machine learning and has been studied as a means of reducing the complexity of algorithms and, more recently, filtering noisy labels in data. The memory and computation requirements of instance-based learning algorithms, such as the nearest neighbor algorithm [1], are proportional to the number of instances in the training set; as a result, instance pruning is often a crucial step in their practical implementation. Recently a number of additional instance pruning approaches have been developed for the task of learning with imperfect labels in classification problems. This is a problem of growing importance in machine learning, due in large part to the advent of crowd-sourcing labeling methods that make attaining labels from a large number of "untrusted" annotators relatively inexpensive. There exist a number of approaches to instance pruning in the classification literature related to learning from imperfect labels, such as evaluation of an annotators accuracy without ground truth labels [2, 3], identifying instances that need relabeling [4, 5].

In this paper we aim to develop a similar method for regression. Most popular statistics methods operate on the assumption that the errors are zero-mean, uncorrelated random variables of equal variance. When these assumptions are violated, in the form of extreme outliers or leverage points, their performance deteriorates. Some possible solutions to this problem are: (a) utilize either regression diagnostics which seek to identify and remove the outliers [6], or (b) use robust regression procedures, which seek to weights instances based on the degree to which they adhere to the underlying parametric model [7, 8]. While a number of robust approaches have been proposed for minimizing the affects of outliers, they still rely on the assumption that the majority of the data adhere to a known parametric model.

This paper introduces a non-parametric exemplar rejection algorithm using Euclidean minimum spanning trees. In particular, we derive a bound on the regression error that can be estimated from the available training data. Using this bound as a starting point, we construct an iterative instance removal algorithm that identifies specific samples with the greatest contribution to this bound. We use the proposed algorithm in a real-world application where we objectively evaluate the perceptual attributes of pathological speech signals. Our approach removes examples with noisy responses during training. Hence it results in a higher performance on a regression task when compared to the two competing alternatives - instance removal approaches [9, 10] and robust regression by soft weighting of outlier examples [8, 11].

The rest of this paper is outlined as follows: Section 2 provides the derivation of an empirically estimable graph-theoretic estimator for the mean squared error of the optimal estimator. Using the results of Section 2 as motivation, Section 3 introduces an exemplar rejection algorithm that will incrementally remove instances in order to approximately minimize the derived MSE estimate. Section 4 demonstrates the performance of the proposed algorithm using synthetic data, and using a real-world example that attempts to model the severity of individuals with speech disorders using the subjective evaluations of professional SLPs.

2. THEORY

The primary motivation for this paper is the scenario where the responses in a regression problem are imperfect. Given a set of data (x_i, y_i) for $i \in [1 \dots n]$, where each instance $x_i \in \mathbf{R}^d$ is sampled from the underlying distribution $f(x)$ and each response $y_i \in \mathbf{R}^1$ is defined as:

$$y_i = \theta(x_i) + \beta_i \tag{1}$$

where $\theta(x_i)$ reflects an oracle function that provides the ground truth response for each exemplar and β_i is zero mean white noise with variance σ_i^2. We assume that both the exemplars and the corresponding responses are independent identically distributed random variables. In this section we will introduce a graph-theoretic measure that upper bounds the mean squared error (MSE) of the optimal estimator of the true response $\theta(x)$. Additionally, we will show that as the number of samples approaches infinity, the proposed function converges to the MSE. To begin, we define a spanning tree on $\mathbf{X}$ as a connected graph, $G = (\mathbf{X}, E)$, with vertex set $\mathbf{X}$, edges E, edge weights given by Euclidean distances between vertices, and no cycles. The length $\mathcal{L}(G)$ of the spanning tree is the sum of its edge lengths. The minimal spanning tree (MST), is defined as the spanning tree with the minimum length. Throughout the rest of this paper we will use G to refer to the MST of $\mathbf{X}$. If each instance x_i from $\mathbf{X}$ is assigned response y_i, we can compute the total squared difference of neighboring responses in the MST,

$$\mathcal{C}(\mathbf{X}, \mathbf{y}) = \sum_{e_{ij} \in E} |y_i - y_j|^2. \tag{2}$$

For every edge in the Euclidean MST of $\mathbf{X}$, this statistic compares the difference in y values between neighboring nodes in G. It is

*School of ECEE and Dept. of SHS, ASU
†IBM Thomas J. Watson Research Center
‡ASU SenSIP Center

easy to show that this statistic is generalization of the well-known Friedman-Rafsky test statistic [12]. In fact, if we consider the binary classification case where $y \in [0, 1]$, the statistic in (2) is equivalent to the Friedman Rafsky test statistic.

Here we show that the expected value of this statistic asymptotically converges to the variance of β_i. Substitute (1) into (2) we get:

$$\begin{aligned} E\big[\mathcal{C}(\mathbf{X}, \mathbf{y})\big] &= E\left[\sum_{e_{ij}\in E} |\theta(x_i) + \beta_i - \theta(x_j) - \beta_j|^2\right] \\ &= \sum_{e_{ij}\in E} E\big[(\theta(x_i) - \theta(x_j))^2\big] + \sum_{e_{ij}\in E} E\big[(\beta_i - \beta_j)^2\big] \\ &+ \sum_{e_{ij}\in E} 2E\big[(\beta_i - \beta_j)(\theta(x_i) - \theta(x_j))\big] \\ &= \sum_{e_{ij}\in E} E\big[(\theta(x_i) - \theta(x_j))^2\big] + \sum_{e_{ij}\in E} (\sigma_i^2 + \sigma_j^2) \end{aligned} \tag{3}$$

The second term in (3) is the sum of the variances for each set of points connected by an edge in the MST. Without loss of generality, we can simplify this term by defining the degree of each vertex in the MST as ρ_i for $i \in [1 \dots n]$. It is easy to see that this variance term in (3) can be rewritten as

$$\sum_{e_{ij}\in E} (\sigma_i^2 + \sigma_j^2) = \sum_{i=1}^{n} E[\rho_i]\sigma_i^2. \tag{4}$$

This corresponds to weighted sum of the variance of each label. A number of theoretical and empirical studies suggest that the degree distribution in random graphs follows a power law, $f_\rho(\rho) \sim (\frac{1}{\rho})^\gamma$ [13]. This is known to be a relatively low-entropy distribution, where the variable takes one of only a few values with high probability. As a result, the weights in the sum would not exhibit great variability, especially for large values of γ. If we now assume that the noise is identically distributed ($\sigma_i^2 = \sigma^2$ for $i \in [1 \dots n]$), this weighted sum simplifies to

$$\sum_{i=1}^{n} E[\rho_i]\sigma_i^2 = \sigma^2 \sum_{i=1}^{n} E[\rho_i] \tag{5}$$

It is known that the sum of the degrees across all nodes in a spanning tree is equal to two times the number of edges ($2||G||$), and that the number of edges is one less than the number of nodes, therefore

$$\sigma^2 \sum_{i=1}^{n} E[\rho_i] = \sigma^2(2||G||) = 2(n-1)\sigma^2. \tag{6}$$

Now let us define

$$\Phi(\mathbf{X}, \mathbf{y}) = \frac{\mathcal{C}(\mathbf{X}, \mathbf{y})}{2(n-1)}, \tag{7}$$

with mean

$$\begin{aligned} E[\Phi(\mathbf{X}, \mathbf{y})] =& \frac{1}{2(n-1)} \sum_{e_{ij}\in E} E[(\theta(x_i) - \theta(x_j))^2] \\ &+ \frac{1}{2(n-1)} \sum_{i=1}^{n} E[\rho_i]\sigma_i^2. \end{aligned} \tag{8}$$

Under this assumption, it is easy to see that the first term in (8) will be greater than or equal to zero, therefore Φ upper bounds a weighted version of the MSE of the ideal estimator. Furthermore, if $\theta(x)$ is a Lipschitz continuous function with Lipschitz constant K, then $\|\theta(x_i) - \theta(x_j)\|^2 \leq K\|x_i - x_j\|^2$. Combining this relationship with the assumption that the noise is identically distributed in (5) and (6), we can form the following inequality:

$$E[\Phi(\mathbf{X}, \mathbf{y})] \leq \frac{K}{2(n-1)} \sum_{e_{ij}\in E} \|x_i - x_j\|^2 + \sigma^2. \tag{9}$$

This not only bounds the MSE of the ideal estimator, we will go on to show that when x is drawn from a compactly supported distribution $f(x)$, it asymptotically converges to the true noise variance, σ^2, since the first term in the inequality goes to 0. If we define the average euclidean distance of the edges in G as

$$\Gamma(G) = \frac{1}{n-1} \sum_{e_{ij}\in E} \|x_i - x_j\|^2, \tag{10}$$

then the first term in (9) is equal to $\frac{K}{2}\Gamma(G)$. This term manages the tightness of the MSE bound. This quantity has been studied in the literature. In fact, in [14] Steele showed that, with probability 1,

$$\lim_{n\to\infty} \Gamma(G) = \lim_{n\to\infty} c(d)(n-1)^{\frac{-2}{d}} \int_{\Re} f(x)^{\frac{d-2}{d}} dx, \tag{11}$$

where $f(x)$ represents the compactly supported probability density function of the underlying distribution from which $\mathbf{X}$ is sampled, $c(d)$ denotes a strictly positive constant, and d represents the data dimension. Note that this expression is only valid when $d > 2$.

We see from (11) that the bias of Φ is affected by 1) the Lipschitz constant; 2) the number of samples; 3) the dimension; 4) the density of the input distribution. From this relationship it is easy to see that the bound becomes arbitrarily tight as $n \to \infty$. Furthermore, the rate of convergence for smaller d, smaller K, and more compact distributions, $f(x)$, increases.

3. ALGORITHM

The overall goal of this paper is to investigate methods of identifying subsets of data containing noisy responses so that they can be removed before training the model. A starting point for the algorithm is the theory outlined in the previous section. Because $\Phi(\mathbf{X})$ asymptotically converges to the label variance, we propose an algorithm which seeks to identify the subset of instances Ω that minimizes the following metric w.r.t. Ω:

$$\Phi(\mathbf{X}(\Omega), \mathbf{y}) = \frac{1}{(2|\Omega| - 1)} \sum_{e_{ij}\in E} |y_i - y_j|^2 \tag{12}$$

There are two problems with seeking to minimize this criterion. First, evaluating this criterion for every potential subset of exemplars would require construction of $2^n - 1$ minimum spanning trees, which will be computationally infeasible even for relatively small data sets. Second, minimizing this criterion without penalizing or restricting the number of exemplars removed would yield a subset containing the two instances with the closest y values. We remedy both of these problems by using a sequential backward selection (SBS) algorithm which will iteratively remove the "worst" instances until it reaches the desired subset size k. By using the SBS algorithm we reduce the required number of MSTs to $\sum_{n=k+1}^{N} n = \frac{1}{2}(N-k)(N+k+1)$. For large values of N, this approach may still be computationally

Algorithm 1 Iterative exemplar removal using minimum spanning trees

Input: Data Matrix $\mathbf{X}$, Stopping Criteria k
Output: Top $N-k$ exemplars that minimize Φ : Ω
Define: $\Omega = 1 \dots N$
for $j \in 1 \dots k$ **do**
  $\mathcal{G}(E, \mathbf{X}(\Omega)) = MST(\mathbf{X}(\Omega))$
  for $i \in \mathcal{G}(E, \mathbf{X}(\Omega))$ **do**

$$\Psi_i(\mathbf{X}(\Omega), \mathbf{y}) = \frac{1}{\rho_i} \sum_{e_{ij} \in E} |y_i - y_j|^2$$

  end for
  $\Omega = \Omega \setminus \underset{i}{\mathbf{argmax}}\, \Psi_i$
end for

prohibitive in many applications. To further reduce the computational burden we introduce an alternate criterion that represents average squared difference in labels across branches connected to point x_i

$$\Psi_i(\mathbf{X}, \mathbf{y}) = \Phi(\mathbf{X}, \mathbf{y}|i) = \frac{1}{\rho_i} \sum_{e_{ij} \in E} |y_i - y_j|^2 \quad (13)$$

In short, this function estimates the average difference between y values for all points connected to x_i in the MST. As such, it estimates the contribution of instance x_i to $\Phi(\mathbf{X}(\Omega))$. While this is a heuristic simplification of the optimal criterion in (12), empirical simulations found that the new, simplified criterion performs well on a number of experiments (see Section 4). By using the Ψ criterion, the algorithm now requires the construction of k MSTs.

4. RESULTS

This Section is broken into two parts. Section 4.1 introduces an academic example where synthetic data are corrupted by various levels of noise and the proposed algorithm is used to remove the noisy instances. Section 4.2 attempts to apply the proposed algorithm in real data to identify and remove instances that do not fit with their subjective evaluation.

4.1. Academic Simulation

To test the effectiveness of the proposed algorithm we examine the scenario where y can be expressed as a linear combination of x plus additive gaussian white noise (AGWN) $\beta \sim N(0, \sigma^2)$.

$$y_i = \sum_{j=1}^{M} \alpha(j) x_i(j) + \beta_i = \tilde{y}_i + \beta_i \quad (14)$$

where M represents the number of features and $x_i(j)$ represents the value of feature j for instance i. For our simulation, we make X a 25-dimensional matrix containing 2000 exemplars that are normally distributed with zero mean and unit variance. We set all α coefficients equal to one while β may take on one of 5 different noise levels with equal probability.

Once the dataset is generated, we use Alg. 1 to iteratively remove exemplars from the dataset. After each exemplar is removed, we use the reduced dataset to train a linear model. We evaluate the performance of each model based on its mean squared error (MSE) in predicting the true labels $\tilde{y}_i$. The resulting MSE values are displayed in Figure 2. It is clear from this figure that removing the high-noise values results in a smaller MSE. In fact, there is an initial dramatic drop in the the MSE (resulting from removing the highest-noise exemplars), followed by a more gradual decline. To verify that the algorithm is correctly identifying the exemplars generated from the higher σ^2 values we also plot the histogram displaying how many of the first 400 exemplars removed are drawn from each noise level in Figure 1. This Figure shows that while the algorithm does select instances for rejection from every noise level, the majority of exemplars that are removed are drawn from the higher noise levels.

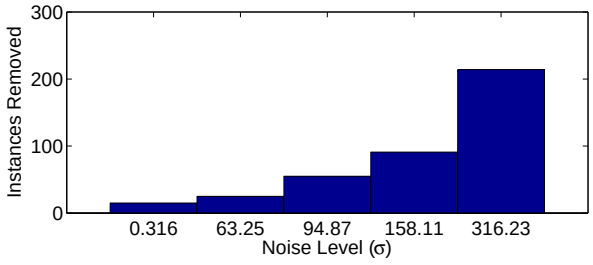

Fig. 1. Number of instances, among the first 400 removed by the proposed algorithm, corresponding to each noise level

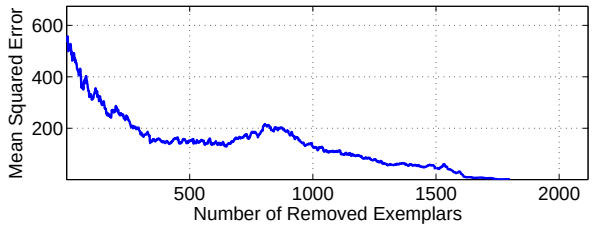

Fig. 2. MSE as a function of the number of exemplars removed.

4.2. Disordered Speech Example

The target application for the work in this paper is the objective evaluation of the perceptual attributes of pathological speech. While there exist a number of different perceptual attributes to investigate, we use severity as a test case for the analysis presented in this paper.

4.2.1. Data

We make use of data collected in the Motor Speech Disorders Laboratory at Arizona State University, consisting of 34 dysarthric speakers and 13 healthy speakers. The dysarthria speakers included: 12 speakers with ataxic dysarthria, secondary to cerebellar degeneration, 10 mixed flaccid-spastic dysarthria, secondary to amyotrophic lateral sclerosis, 8 speakers with hypokinetic dysarthria secondary to Parkinson's Disease, and 4 speakers with hyperkinetic dysarthria secondary to Huntington's disease (HD). Each patient provided speech samples, including a reading passage, phrases, and sentences. The speech database consists of approximately 10 minutes of recorded material per speaker. In addition to identifying each speaker based on their specific neurogenic disorder, the database provides subjective evaluations for each speaker from 6 different speech-language pathologists. The 6 SLPs were blinded to the speakers dysarthria subtype, and were asked to mark each speaker along 5 different perceptual attributes (severity, nasality,vocal quality, articulatory precision, prosody) on a scale ranging from normal to severely abnormal. A number of candidate feature sets have been developed

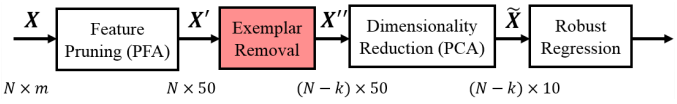

Fig. 3. Block diagram of experimental design.

in the literature for analysis of pathological speech [15–20]. In this paper, we restrict our analysis to the long-term average spectrum (LTAS) [21], Envelope Modulation Spectrum [22], and P.563 [23]. This results in a total of 4180 instances each represented by 123 features. A detailed description of each feature set can be found in [17]. These speech samples were taken from the larger pathological speech databased described in [24].

4.2.2. *Experimental Setup*

A graphical depiction of the experimental setup is provided in Figure 3. Beginning with the feature 4180x123 feature matrix described in 4.2.1, we partition the data for cross-validation by removing a single speaker from the training set and placing them in the test set. We then employ Principal Feature Analysis algorithm [25] with the correlation matrix to identify the top 50 features with minimal redundancy. We then apply Principal Components Analysis to further reduce the dimensionality. The instance rejection procedure described in Alg. 1 is applied to the resulting dataset to iteratively identify and remove noisy instances. We set $K = 1000$ and select 5 points per MST for removal, and monitor the performance of the algorithm as a function of the number of exemplars removed. After removing the noisy instances we use PCA to form a low-dimensional representation. Finally we use the modified dataset, which now contains $N - k$ exemplars expressed in 10 dimensions, to train a linear model using robust least squares with Huber loss. The severity ratings of a single SLP are used as the response variable. The resulting model is then used to predict responses for each of the instances that were partitioned into the test dataset. This yields a set of responses pertaining to a single speaker that we then average to achieve an estimate of the severity rating for that speaker. This process is then repeated until we have generated severity ratings for all 33 speakers and 6 SLPs in the data set.

In addition to the exemplar selection method proposed in this paper, we also tested 3 alternate selection methods for selecting examples: (a) k-means clustering, (b) support vector regression, and (c) speaker based exemplar removal. In k-means exemplar selection we identify 250 clusters using the k-means algorithms, then iteratively remove points with the largest distance from their respective cluster mean. Next we a construct a support vector machine (SVM), and remove all instances not chosen as support vectors. We control the number of support vectors by varying the C parameter from 0.025 to 10 on a logarithmic scale while ϵ is held constant at 0.25. Finally we employ a speaker based pruning method where we look at the data from each individual speaker and remove iteratively remove the instances with the largest standardized euclidean distance from the mean.

4.2.3. *Evaluation*

Because severity is inherently subjective and no ground truth scores exist, we use the average severity rating of the 5 SLPs not used in training to approximate the true severity rating. We evaluate the performance of each model by correlating the predicted ratings with the average rating of the 5 SLPs not used to train the model. These correlation values are then averaged across the 6 different SLP CV

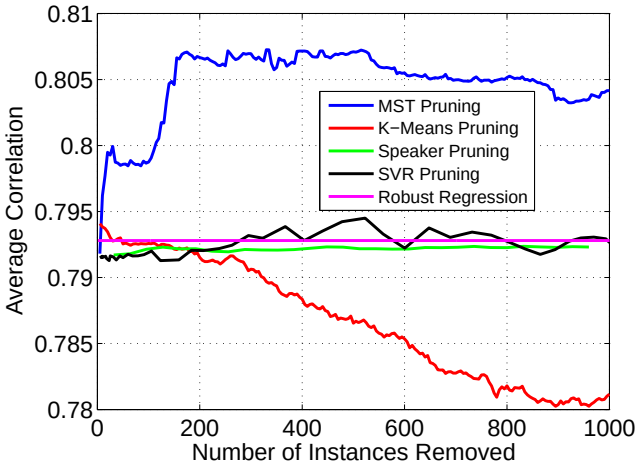

Fig. 4. Average correlation between model predictions and the average subjective rating as a function of the number of exemplars removed. Mean correlation of individual SLPs with the average score is 0.9332)

stages and the resulting average correlation scores are displayed as a function of the number of exemplars removed in Figure 4. Figure 4 shows that of the four exemplar pruning methods that were tested, the proposed MST-based pruning approach is the only one that achieved a noticeable improvement in correlation when averaged across the six SLPs. The proposed method yields an improvement of ~ 0.015 in the average correlation score which plateaus at around 150 instances and begins to decline after $\sim$ 500 exemplars have been removed. This decline marks the point at which the benefit of reducing the noise in the dataset no longer outweighs the information lost in reducing the overall size of the dataset.

5. CONCLUSION

In real world regression problems, classical regression algorithms must be employed cautiously due to the strict assumptions under which they successfully operate. In this paper we discuss a new method for bounding the minimal error in regression problems using a graph-theoretic approach. Furthermore, this bound is non-parametric and can be empirically estimated given only samples and corresponding responses. Motivated by this bound, we present an iterative algorithm for data removal. The algorithm aims to minimize the new bound by selecting points with the greatest contribution to the bound. We evaluate the algorithm on a pathological speech assessment task, where we aim to learn a model that predicts expert scores of different atypical speech samples. Results show that, by removing the noisy points, we obtain higher quality models that better correlate with expert perception.

6. ACKNOWLEDGMENTS

This research was supported in part by the Office of Naval Research grant N000141410722 (Berisha) and National Institute of Health, National Institute on Deafness and Other Communicative Disorders Grant (Liss and Berisha)

7. REFERENCES

[1] Thomas Cover and Peter Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[2] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*. ACM, 2009, pp. 889–896.

[3] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[4] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 614–622.

[5] Carla E Brodley and Mark A Friedl, "Identifying mislabeled training data," *arXiv preprint arXiv:1106.0219*, 2011.

[6] Peter J Rousseeuw and Annick M Leroy, *Robust regression and outlier detection*, vol. 589, John Wiley & Sons, 2005.

[7] Paul W Holland and Roy E Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-Theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[8] Peter J Huber, *Robust statistics*, Springer, 2011.

[9] Stuart Lloyd, "Least squares quantization in pcm," *Information Theory, IEEE Transactions on*, vol. 28, no. 2, pp. 129–137, 1982.

[10] Alex J Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[11] John A Hartigan and Manchek A Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[12] Jerome H Friedman and Lawrence C Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *The Annals of Statistics*, pp. 697–717, 1979.

[13] PJ Macdonald, E Almaas, and A-L Barabási, "Minimum spanning trees of weighted scale-free networks," *EPL (Europhysics Letters)*, vol. 72, no. 2, pp. 308, 2005.

[14] J Michael Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *The Annals of Probability*, pp. 1767–1787, 1988.

[15] Visar Berisha, Rene Utianski, and Julie Liss, "Towards a clinical tool for automatic intelligibility assessment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2825–2828.

[16] Visar Berisha, Steven Sandoval, Rene Utianski, Julie Liss, and Andreas Spanias, "Selecting disorder-specific features for speech pathology fingerprinting," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7562–7566.

[17] Visar Berisha, Julie Liss, Steven Sandoval, Rene Utianski, and Andreas Spanias, "Modeling pathological speech perception from data with similarity labels," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 915–919.

[18] Steven Sandoval, Visar Berisha, Rene L Utianski, Julie M Liss, and Andreas Spanias, "Automatic assessment of vowel space area," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL477–EL483, 2013.

[19] Visar Berisha, Steven Sandoval, Rene Utianski, Julie Liss, and Andreas Spanias, "Characterizing the distribution of the quadrilateral vowel space area," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 421–427, 2014.

[20] Alan Wisler, Visar Berisha, Julie Liss, and Andreas Spanias, "Domain invariant speech features based on a new divergence measure," in *Proceedings of IEEE Spoken Language Technology Workshop*. IEEE, 2014.

[21] Phil Rose, *Forensic speaker identification*, CRC Press, 2003.

[22] Julie M Liss, Sue LeGendre, and Andrew J Lotto, "Discriminating dysarthria type from envelope modulation spectra," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, 2010.

[23] ITUT Rec, "P. 563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," *International Telecommunication Union, Geneva*, 2004.

[24] Kaitlin L Lansford and Julie M Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 1, pp. 57–67, 2014.

[25] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 301–304.