
Classification on incomplete data using sparse representations: Imputation is optional

Jort Gemmeke

Dept. of Linguistics, Radboud University,
P.O. Box 9103, NL-6500 HD, Nijmegen, The Netherlands

J.GEMMEKE@LET.RU.NL

Abstract

We present a non-parametric technique capable of performing classification directly on incomplete data, optionally performing imputation. The technique works by sparsely representing the available data in a basis of example data. Experiments on a spoken digit classification task show significant improvement over a baseline missing-data classifier.

1. Introduction

Classification on incomplete data is a challenging task because parametric techniques require that the dimensionality of the data doesn't change between training and classification while non-parametric techniques which can handle incomplete data such as k-nearest-neighbors often deliver suboptimal classification accuracies. In practice, the missing data is often estimated prior to classification through *imputation*. Most imputation methods estimate the missing coefficients based on local information and/or do not fully exploit the structure of the underlying signal. Based on work in *Compressed Sensing* (Donoho, 2006; Candes, 2006) we present a non-parametric method which can not only perform classification directly on the available data but optionally imputes the missing data. The method is based on the premise that a signal can be sparsely represented in a basis of example signals (Yang et al., 2007) and that this sparse representation can be exactly recovered even if only a small part of the data is available (Zhang, 2006). We show the effectiveness of this approach on a spoken digit classification task.

2. Method

2.1. Sparse representation

We consider observation vector \mathbf{y} of unknown class and dimensionality K to be a linear combination of feature vectors $\mathbf{d}_{i,n}$, where the first index ($1 \leq i \leq I$) denotes

one of I classes and the second index ($1 \leq n \leq N_i$) a specific exemplar vector of class i with N_i the number of examples in that class. We write:

$$\mathbf{y} = \sum_{i=1}^I \sum_{n=1}^{N_i} \alpha_{i,n} \mathbf{d}_{i,n}$$

with weights $\alpha_{i,n} \in \mathbb{R}$. The set of exemplars span a $K \times N$ dimensional basis ($N = N_1 + N_2 + \dots + N_I$):

$$A = (d_{1,1} \dots d_{1,N_1} \dots d_{I,1} \dots d_{I,N_I}) \quad (1)$$

Thus, we can express \mathbf{y} as:

$$\mathbf{y} = A\mathbf{x} \quad (2)$$

with \mathbf{x} an N -dimensional vector that ideally will be sparsely represented as $\mathbf{x} = [0 \dots 0 \alpha_{i,1} \alpha_{i,2} \dots \alpha_{i,N_i} 0 \dots 0]^T$ (i.e., most coefficients not associated with class i are zero).

Taking into account that the observation vector \mathbf{y} is incomplete we denote its available coefficients by \mathbf{y}_a and the missing coefficients \mathbf{y}_m . Now we can solve the system of linear equations of Eq. 2 using only the available coefficients \mathbf{y}_a and the basis A_a , formed by only retaining the rows of A indicated by the available coefficients. Research in the field of *compressed sensing* (Donoho, 2006; Candes, 2006) has shown that if \mathbf{x} is sparse, \mathbf{x} can be recovered exactly by solving:

$$\min \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{y}_a - A_a \mathbf{x}\|_2 \leq \epsilon \quad (3)$$

with a small constant ϵ such that the error \mathbf{e} satisfies $\|\mathbf{e}\|_2 < \epsilon$ and $\|\cdot\|_1$ the l^1 norm.

2.2. Sparse classification (SC)

Following (Yang et al., 2007), we perform classification by comparing the *support* of \mathbf{y}_a in parts of A_a associated with different classes i . In other words, we compare how well the various parts of \mathbf{x} associated with different classes i can reproduce \mathbf{y}_a . The reproduction error is called the *residual*. The residual of

class i is calculated by setting the coefficients of \mathbf{x} not associated with i to zero while keeping the coefficients associated with i unchanged. Thus the residual is:

$$r_i(\mathbf{y}_r) = \|\mathbf{y}_a - A_a \delta_i(\mathbf{x})\|_2 \quad (4)$$

with $\delta_i(\mathbf{x})$, the vector selecting only the columns of A that correspond to class i . The class c that is assigned to an observed vector \mathbf{y} is the one that gives rise to the smallest residual:

$$c = \underset{i}{\operatorname{argmin}} r_i(\mathbf{y}_r). \quad (5)$$

2.3. Sparse imputation (SI)

Alternatively, one can use the sparse representation \mathbf{x} to impute the missing coefficients. Without loss of generality we reorder \mathbf{y} and A as in (Zhang, 2006) so that we can write:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{y}_a \\ A_m \mathbf{x} \end{bmatrix} \quad (6)$$

with A_m pertaining to the rows of A indicated by the missing coefficients in \mathbf{y} and \mathbf{y}_i an estimate for the missing coefficients \mathbf{y}_m . This yields a new observation vector $\hat{\mathbf{y}}$ after which ordering can be restored.

3. Experiments

We apply the described method to missing data spoken digit classification task (AURORA-2). In noisy speech, with digits represented by fixed length observation vectors, coefficients are considered missing if their values (representing speech energy in a time-windowed spectrographic representation) are dominated by speech energy rather than noise energy. We explore the effectiveness of our approach by selecting the missing coefficients using knowledge of the true speech and noise signals. Using a setup described in detail in (Gemmeke & Cranen, 2008) we compare the sparse classification technique with a baseline, state-of-the-art, HMM-classifier which performs imputation on a frame-by-frame basis (Van hamme, 2006). Additionally we compare classification accuracies obtained by combining *sparse imputation* and the baseline classifier.

While not strictly linear as a function of signal-to-noise ratio (SNR), the percentage missing coefficients ranges from 0% (clean speech) to 80 – 95% (at SNR -5 dB). In Table 1 it is shown that the *sparse classification* and *sparse imputation* methods significantly outperform the baseline, frame-based classifier.

Table 1. AURORA-2 single digit classification accuracy.

method	SNR					
	clean	15	10	5	0	-5
Baseline	99.3	99.1	98.7	96.6	88.4	61.0
SC	98.4	98.4	98.0	97.5	95.8	91.0
SI	99.3	99.0	98.5	97.7	96.5	91.3

4. Discussion and conclusions

Results show that both sparse methods give considerable improvement over the baseline, suggesting that a correct sparse representation can be found even when the majority of the data is missing, provided the redundancy in the structure of the data is exploited by use of example whole-digit observations vectors. The slightly better results using sparse imputation rather than sparse classifications seem to suggest that the sparse classification method does not generalize to observed digits as well as the HMM-based (parametric) approach.

Acknowledgments

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

References

- Candes, E. J. (2006). Compressive sampling. *Proceedings of the International Congress of Mathematicians*.
- Donoho, D. L. (2006). For most large underdetermined systems of equations, the minimal ℓ_1 -norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59, 907–934.
- Gemmeke, J., & Cranen, B. (2008). Noise robust digit recognition using sparse representations. *accepted for ISCA ITWR 2008*.
- Van hamme, H. (2006). Handling time-derivative features in a missing data framework for robust automatic speech recognition. *Proceedings of IEEE ICASSP*.
- Yang, A. Y., Wright, J., Ma, Y., & Sastry, S. S. (2007). Feature selection in face recognition: A sparse representation perspective. *submitted to IEEE Transactions Pattern Analysis and Machine Intelligence*.
- Zhang, Y. (2006). When is missing data recoverable? *Technical Report*.