USING COMPRESSED AUDIO-VISUAL WORDS FOR MULTI-MODAL SCENE CLASSIFICATION

Jan J. Kurcius and Toby P. Breckon*[+]

*Cranfield University, Bedfordshire, UK   [+]Durham University, Durham, UK

ABSTRACT

We present a novel approach to scene classification using combined audio signal and video image features and compare this methodology to scene classification results using each modality in isolation. Each modality is represented using summary features, namely Mel-frequency Cepstral Coefficients (audio) and Scale Invariant Feature Transform (SIFT) (video) within a multi-resolution bag-of-features model. Uniquely, we extend the classical bag-of-words approach over both audio and video feature spaces, whereby we introduce the concept of compressive sensing as a novel methodology for multi-modal fusion via audio-visual feature dimensionality reduction. We perform evaluation over a range of environments showing performance that is both comparable to the state of the art (86%, over ten scene classes) and invariant to a ten-fold dimensionality reduction within the audio-visual feature space using our compressive representation approach.

Index Terms— multi-resolution, bag of words, MFCC, compressed sensing, audio-visual, multi-modal

1. INTRODUCTION

Automating the process of scene understanding and classification is a significantly large challenge [1][2]. Many prior works have attempted to replicate human abilities for this task, such that actions dependent upon scene understanding can maximise the chances of achieving a given goal. Notably, machine perception tasks are often considered in a single modality, whereas we as humans rarely rely on a single sensing modality, performing the same tasks within a contextual bias [3][4]. In addition to emulating human abilities our work is also motivated by the fact that the most readily available mobile devices today, have both audio and visual signal capture capabilities. As a result, employing the second available source of sensing seems a viable solution in many practical situations.

Prior work on scene classification uses a range of both audio [2][5], visual [1] and multi-modal feature classification [6][7]. An existing approach to environment and event classification presented by [2] is purely based on audio signal. Here the authors employ the Mel-frequency Cepstral Coefficients (MFCC) for audio description, performing classification over ten environment classes *{office, lecture, bus, urban driving, railway station, beach, bar, laundrette, football match, city centre street}*. In this work [2] a Hidden Markov Model trained with a Viterbi algorithm achieves a per class accuracy between 75% - 100% on this ten class task.

The recent visual scene classification work of [1] proposes the highly successful multi-resolution bag-of-features model. This improves significantly over the classical bag-of-words approaches for scene classification [8][9] achieving an accuracy of ~84%. Notably, earlier work does not take into account a spatial layout of features [10] [11]. The approach of [1] overcomes this weakness by extracting visual features from a spatial layout of horizontal and vertical image partitions, concatenating the resulting feature codewords within a multi-resolution framework and achieving classification via a Support Vector Machine (SVM). By contrast, scene classification work using combined audio and visual features is in its infancy [5][6] [7].

In prior audio-visual classification work a notable task is that of emotion classification [6]. In this work [6] the model proposed is responsible for assigning a set of audio-visual features to a number of emotion classes with the unsolved problem identified as feature selection. The overall accuracy achieved for this method was ~85%. Among other successful audio-visual classification is a framework presented by Bicego *et. al.* [7]. Their study focuses on event recognition for surveillance purposes, achieving very satisfactory results of ~89%. Studies developed in this area are presented in [5], where authors survey the recent advances in audio-visual affect recognition [12][13].

In contrast to earlier work upon this topic [5][6][7] ,we propose an approach combining the use of multi-resolution bag-of-visual-words [1] and the highly successful MFCC audio features [2] in a novel multi-modal classification approach. Recognising the sparseness of this classification feature space [1][2], we propose a novel methodology that utilises the representative power of compressive sensing [14] to facilitate the derivation of *truly fused* audio-visual words over which we then apply both SVM and Decision Forest [15] classification techniques. As the goal of our approach is general and unconstrained audio-visual scene classification, it is different from earlier feature concatenation techniques, such as the Audio-Video Concurrence (AVC) utilised in [7] or Affective Audio-Visual Words employed in [6]. In contrast to these and earlier feature concatenation based approaches we show that our compressive audio-visual feature representation facilitates a significant reduction in dimensionality with only marginal impact on the resulting classification performance.

2. AUDIO-VISUAL WORDS

Within our proposed framework we firstly outline our audio and visual feature representations.

2.1 Visual Word Representation

Our visual words are extracted from images using the seminal Scale-Invariant Feature Transform (SIFT) [16] to obtain feature descriptors applied within the multi-resolution bag-of-features technique of [1]. All images are pre-processed, prior to the SIFT feature extraction using the real-time saliency detection approach of [17]. This reduces overall scene complexity and aids the later derivation of a sparse visual scene representation.

In order to generate visual scene descriptors we first generate a vocabulary of visual codewords following the classical bag-of-visual-features approach. This is performed by first, sub-sampling each training image, over a given set, by a factor of two and performing dense SIFT feature extraction over an image pyramid (*depth = 3, sampling grid: 40x30*) [1]. In this work we use a training set of 1300 images to generate this fixed length vocabulary of codewords via *k*-means clustering [18].

Based on this visual vocabulary we then generate visual scene descriptors, based on SIFT features detected in horizontal and vertical partitions across the image following [1] (illustrated in Fig. 1). A visual codeword is created redundantly for each vertical and horizontal partition over multiple scales, s (s = 3). For each level in this pyramid, the original image is partitioned into eight horizontal and vertical subregions for s=*1*, four subregions for s=*2* and two subregions for s=*3*, thus producing 28 image partitions in total (Fig. 1.). The visual codewords extracted from all of these partitions are then concatenated to produce our global *28k* dimension multi-resolution visual descriptor (Fig. 1).

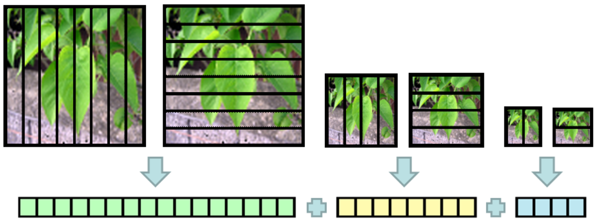

Fig. 1. Contribution of each resolution image to the image representation

In this work we consider the use of hard-assignment in visual word construction as it results most frequently in a sparse signal distribution within the codeword representation and this is known to be most suited to the compressive sensing derived feature combination approach which we employ (Section 2.3). Notably, our approach to visual descriptor construction differs from the original work of [1] in terms of classification methodology applied. Here, we intend to use global classification upon this image descriptor as opposed to independent classification, performed at each of pyramid resolution, as per [1].

2.2. Audio Word Representation

We represent the audio signal utilising the widely used Mel-frequency Cepstral Coefficients (MFCC) [19]. MFCC extraction is comprised over a number of basic steps. Firstly, the audio signal is divided into small, overlapping frames, with the approximate frame length of 20ms, allowing consideration of them as a periodic signal. Furthermore, in order to avoid spectral leakage caused via discontinuities at each end of the frame, we multiply an audio signal by a widely used Hamming window function [20]. Secondly, we compute the power spectrum of each windowed frame which results in short-time Fourier transform [21]. Subsequently, each frame is filtered in a Mel-frequency domain using a set of triangular, overlapping filters with a variable width increasing with frequency scale and thus reflecting human audio sensitivity [22]. Following this operation, the resulting energy values for frequencies within the same Mel-filterbank are summed, resulting in a set of n values, where n is a number of Mel-filterbanks. Subsequently, we calculate the discrete cosine transform over this set, treating it as a signal. Finally, the amplitude of the resulting spectrum creates a set of desired MFCC [23], equal in size to the number of Mel-filterbanks used. Our representation is based on the use of a significant training set of audio samples over which the MFCC sets associated to each extracted frame are clustered using *k*-means. The set of resulting clusters is then used as an audio-word vocabulary, akin to the visual-word vocabulary of Section 2.1, for describing unseen MFCC examples derived from frames associated with single audio sample.

2.3. Multi-modal Feature Representation

The most significant challenge in multi-modal classification is a derivation an effective feature combination or fusion methodology. The simple concatenation of multi-modal feature representations is a commonplace [5][6]. By contrast, here we look to the use of a compressive sensing derived methodology [14][24] as an approach for the combinatorial mapping of a multi-modal feature space into a single compressed multi-dimensional representation. Prior work in compressive sensing [14] has shown that such a combination can be formed with minimal loss of the underlying information contained within the larger signal and this differs significantly from other dimensionality reduction schemes such as PCA or LDA [14].

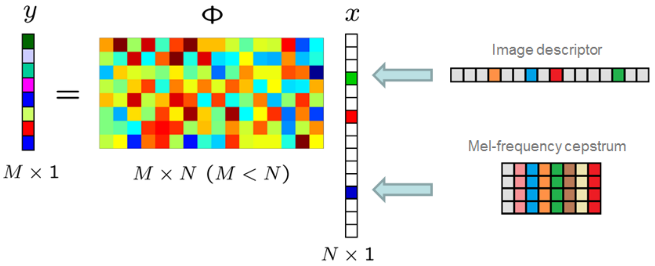

Fig. 2. Multi-modal fusion via compressive sensing projection

The motivation for this approach comes from the fact that the resulting codeword representation in both modalities is significantly sparse. A compressive sensing methodology provides a solution for this issue, by multiplying the concatenated representations of image and audio training data samples by a random projection matrix [14]. This is illustrated in Fig. 2, where we can see the projection matrix of size *M x N*, where *N* is the length of the audio-visual sample representation, and *M* is the length of the resulting compressed audio-visual feature representation. This results in a compressed vector representation (*y*), which is in significantly lower dimension (*M*) and dense in nature [14].

3. CLASSIFICATION

In this study classification is performed via Support Vector Machine utilising the Radial Basis Function (RBF) kernel [25] and Decision Forests [26]. SVM training is carried out using grid search over the kernel parameter space [27] and similarly Decision Forests are evaluated over a range of parameters (at most 100 trees, max. depth = 25, min. amount of samples per leaf node = 5 and max. categories = 15). The evaluation of each classification approach is determined using random sub-set based cross-validation over a set of 2000 samples (set size = 1000) [28]. Training data is generated from previously prepared audio-visual samples over a set of ten environment classes: *{university (outdoor), university (indoor), canteen, city centre, railway station, motorway, footpath, shopping centre, open-air market, bus}*. Furthermore, we evaluate the use of the standard, uncompressed feature-space for this environment classification task using the concatenated audio-visual descriptors (Section. 2.3) in their original dimension, *N*. Our evaluation is performed both with the proposed joint audio-visual feature representation (Section 2.3) and for comparison each of the uncompressed audio and visual feature representations separately (Sections 2.1 / 2.2).

All data samples have been collected using a *640 x 480* image resolution (MPEG-2 compression). The corresponding, synchronised audio signal is captured as a 16-bit uncompressed data (48 kHz sampling). Based on the data gathered, we generated four pairs of vocabularies with visual vocabularies containing k_v = *{100, 400, 700, 1000}* words and with audio vocabulary size $k_a = 28k_v$ (due to the image partitioning approach of Section 2.1). This results in audio and visual feature descriptors of the same dimension that contribute equally to same amount of vector elements to the concatenated audio-visual feature descriptor (dimensionality = $28k_a + 28k_v = 56k_v$) prior to compression.

4. RESULTS

In our first experiment, we investigate the classification effectiveness for all four codebook pairs generated (k_v = *{100, 400, 700, 1000}*, $k_a = 28k_v$) for audio-visual feature descriptor generation. Audio-visual feature fusion was performed using compressive sensing with *M=3000* as outlined (Section 2.3). For reference we perform the scene classification results using just the visual descriptor (Fig. 3) and audio descriptor (Fig. 4). The results of this are presented in Fig. 3/4 where we can see that classification via the SVM classifier maintains an overall accuracy between 84% to 87%, regardless of codebook size, outperforming the Decision Forest approach. The Decision Forest is clearly better suited to the examples that use shorter codebooks and this observation is present throughout Fig. 3/4. As we can see, the multi-modal scene classification using the combined and compressed audio-visual word descriptor (Fig. 5, *M = 3000*) outperforms visual modality (Fig. 3) using the SVM classifier and performs very similarly to the performance of the uncompressed audio features (if not marginally outperforming them in some cases).

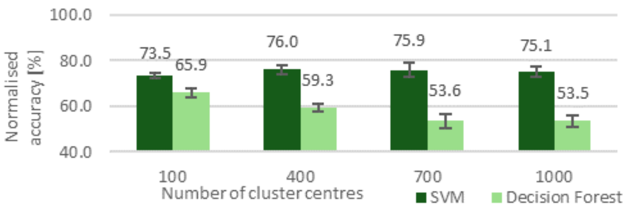

Fig. 3. Scene classification results using visual feature descriptors (uncompressed SIFT feature derived codeword, Section 2.1)

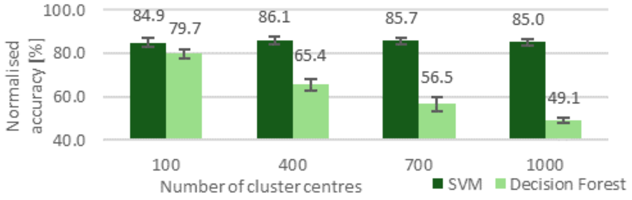

Fig. 4. Scene classification results using audio feature descriptors (uncompressed MFCC feature derived codeword, Section 2.2)

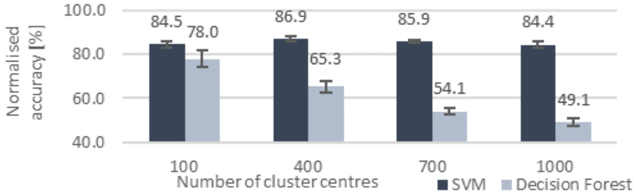

Fig. 5. Scene classification results using audio-visual feature descriptors (compressed audio-visual combined codeword, *M* = 3000)

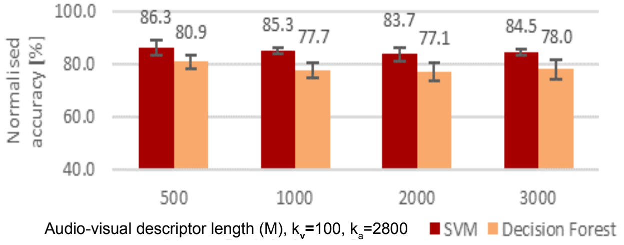

Fig. 6. Classification results obtained using varying levels of combined audio-visual descriptor compression (*M*) (Section 2.3)

From these results we can see that any negative impact on accuracy from using a compressed audio-visual bag-of-words feature representation (Fig. 5) is marginal compared to uncompressed audio (Fig. 4) or non-existent uncompressed visual features (Fig. 5, which actually perform worse). Comparison of the visual feature case with and without visual saliency pre-processing (Section 2.1) resulted in an uniform drop in accuracy (~5%), illustrating the importance of this step within the process.

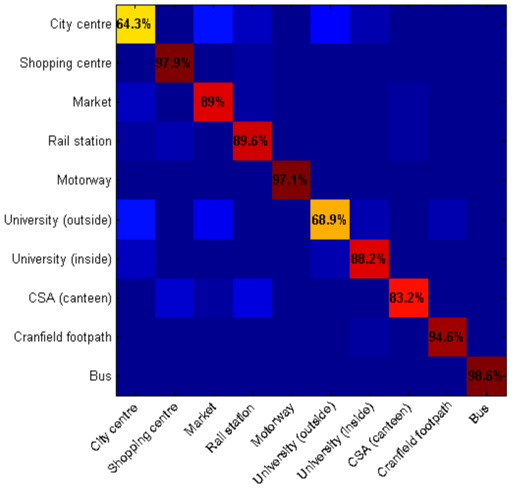

Fig. 7. Confusion matrix for compressed audio-visual descriptor based classification using the SVM classifier ($k_v = 400$; $M = 3000$).

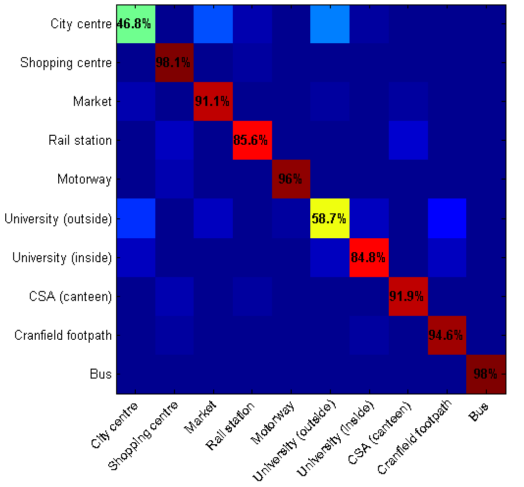

Fig. 8. Confusion matrix for uncompressed audio-visual descriptor classification using the SVM classifier.

Furthermore, we perform a set of classification tests using a variable compression ratio, *M,* of the combined audio-visual descriptor (Fig, 6). As can be seen from the results from both of the classification approaches (Fig. 6), applying higher levels of compression (i.e. $M < 3000$) does not adversely affect overall classification performance. On the contrary, we observe almost the best performance when using the highest compression level (*M=500*, SVM, normalised accuracy = 86.3%). This represents a dimensionality reduction from $56k_v$ $(k_v = 100)$ to $M=500$ (ten-fold) with only a marginal effect on classification performance (comparing Fig. 6 to Fig. 3/4). Comparable results are obtained for $M = 3000$ across a variety of k values (number of clusters). Furthermore we observe a mild increase in classification performance as dimensionality reduction is increased (i.e. *M, 3000 → 500* in Fig. 6). These results (Fig. 3 - Fig. 5) illustrate that the use of a combined yet compressed audio-visual descriptor (Fig. 5) has marginal impact on effective scene classification compared to the use of either visual or audio feature descriptors in isolation. Moreover, the average accuracy achieved when using uncompressed features (84.2%, over all values of k) is 2% lower than the best result achieved using compressed data (86.3%, M = 500, SVM, Fig. 6). The negligible effect on classification observed despite the significant reduction in dimensionality leads to a notable gain in computational performance and bandwidth/storage requirements of classification models (Fig. 3 – Fig. 6).

Additionally, we examine the per class accuracy via the use of confusion matrices presented in Fig. 7 / Fig. 8. Here we can see that the accuracy across the environment classes varies significantly. Using compressed audio-visual descriptors (Fig. 7), we see that all of the classes score above 60% successful detection whereas when using uncompressed audio-visual descriptors have two weak classes (*city centre, university outside*) scoring below 60% (Fig. 8). This indicates the potentially stronger discriminative power of the compressed audio-visual feature representation which impacts upon marginal classification cases. Future work will investigate this in greater detail.

5. CONCLUSIONS

We present a unique approach for audio-visual environment sensing combing state-of-the art methods for both audio and visual feature extraction (multi-resolution bag-of-features model, Mel-frequency Cepstral Coefficients (MFCC) [2] [21]) via a novel multi-modal fusion technique inspired by compressive sensing.

Our evaluation shows marginal impact on classification performance despite a ten-fold reduction in the original audio-visual feature space dimensionality.

Moreover, we have shown that classical bag-of-words approach can be successfully employed for MFCC feature representation in addition to conventional visual descriptors. Our evaluation has shown that this concept produces successful environmental classification with an accuracy comparable to the state of the (~86%) [6][7] that outperforms either of the audio or visual modality in isolation in some cases.

This performance is achieved within a significantly lower dimensional space of audio-visual descriptors and thus at a lower computational cost. Future work will consider a more in-depth investigation of chosen MFCC within a given audio frame as extending MFCC feature set can notably increase classification accuracy [19].

6. REFERENCES

[1] L. Zhou, Z. Zhou, and D. Hu, “Scene classification using a multi-resolution bag-of-features model,” Pattern Recognit., vol. 46, no. 1, pp. 424–433, Jan. 2013.

[2] L. Ma, B. Milner, and D. Smith, “Acoustic environment classification,” ACM Trans. Speech Lang. Process., vol. 3, no. 2, pp. 1–22, Jul. 2006.

[3] K. K. Evans and A. Treisman, “Natural cross-modal mappings between visual and auditory features,” J. Vis., vol. 10, pp. 1–12, 2010.

[4] D. Pascucci and S. Baldassi, “Acoustic cues to visual detection: A classification image study,” J. Vis., vol. 11, no. 6, pp. 1–11, 2011.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: audio, visual, and spontaneous expressions.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[6] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, “Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification,” IEEE Trans. Multimed., vol. 12, no. 6, pp. 523–535, Oct. 2010.

[7] M. Cristani, M. Bicego, and V. Murino, “Audio-Visual Event Recognition in Surveillance Video Sequences,” IEEE Trans. Multimed., vol. 9, no. 2, pp. 257–267, 2007.

[8] P. Perona and L. Fei-Fei, “A Bayesian Hierarchical Model for Learning Natural Scene Categories,” Proc. Conf. Comput. Vis. Pattern Recognit., vol. 2, pp. 524–531, 2005.

[9] L.J. Li and L. Fei-Fei, “What, where and who? Classifying events by scene and object recognition,” Proc. Int. Conf. Comput. Vis., pp. 1–8, 2007.

[10] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] K. Mikolajczyk and C. Schmid, “Performance evaluation of local descriptors.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 10, pp. 1615–30, Oct. 2005.

[12] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernandez-Dols, “Facial and vocal expressions of emotion.,” Annu. Rev. Psychol., vol. 54, pp. 329–49, Jan. 2003.

[13] J. F. Cohn, “Foundations of Human Computing: Facial Expression and Emotion,” Proc. Int. Conf. Multimodal Interfaces, pp. 233–238, 2006.

[14] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.

[15] A. Criminisi, J. Shotton, and E. Konukoglu, “Decision Forests for Classification , Regression , Density Estimation , Manifold Learning and Semi-Supervised Learning,” Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114 5.6 (2011).

[16] D. G. Lowe, “Object recognition from local scale-invariant features,” Proc. Int. Conf. Comput. Vis., pp. 1150–1157 vol.2, 1999.

[17] I. Katramados and T. P. Breckon, “Real-time Visual Saliency by Division of Gaussians,” in Proc. International Conference on Image Processing, 2011, pp. 1741–1744.

[18] M. Halkidi, “On Clustering Validation Techniques,” J. Intell. Inf. Syst., pp. 107–145, 2001.

[19] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, “A Survey of Audio-Based Music Classification and Annotation,” IEEE Trans. Multimed., vol. 13, no. 2, pp. 303–319, Apr. 2011.

[20] G. Heinzel, A. Rudiger, and S. R., “Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top windows”, Technical Report, Max-Planck Institute, Hanover, pp. 1–84, 2002.

[21] E. Jacobsen and R. Lyons, “The sliding DFT,” Signal Process. Mag. IEEE, vol. 20, no. 2, pp. 74–80, 2003.

[22] D. O’Shaughnessy, Speech communication: human and machine. Addison-Wesley Pub. Co., 1987, p. 568.

[23] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust. Speech Signal Process, vol. 28, no. 4, pp. 357–366, 1980.

[24] R. Rigamonti, M. A. Brown, and V. Lepetit, “Are sparse representations really relevant for image classification?,” Proc. Comput. Vis. Pattern Recognition, pp. 1545–1552, 2011.

[25] V. N. Vapnik, Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control Series). Wiley-Interscience, 1998, p. 768.

[26] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[27] C. Chang and C. Lin, “LIBSVM : A Library for Support Vector Machines,” ACM Trans. Intell. Syst. Technol., no. 2:27:1–27:27, pp. 1–39, 2011.

[28] T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.