Affect Sensing in Speech: Studying Fusion of Linguistic and Acoustic Features

Alexander Osherenko   Elisabeth André   Thurid Vogt

University of Augsburg
Germany, Eichleitnerstr. 30, 86159 Augsburg
[osherenko, andre, vogt]@informatik.uni-augsburg.de

Abstract

Recently, there has been considerable interest in the recognition of affect in language. In this paper, we investigate how information fusion using linguistic (lexical, stylometric, deictic) and acoustic information can be utilized for this purpose and present a comprehensive study of fusion. We examine fusion at the decision level and the feature level and discuss obtained results.

1. Introduction

Affect sensing in speech can be used in a wide range of applications, for instance, in dialogue systems or in robotics. However, since natural language is multifold, affect sensing is an error-prone issue. In order to improve classification results, affect sensing can make use of multimodal fusion.

An utterance in a spoken dialogue can be understood both as its text, but also as its acoustic signal. Therefore, affect sensing can be performed by analyzing lexical elements in its text, but also by exploiting acoustic features.

This paper focuses on issues of combining lexical and acoustic features. Hereby, we concentrate on the following questions:

1. Is fusion beneficial for affect sensing at all?
2. Should we consider the context of an utterance to improve affect sensing?
3. What is more beneficial for affect sensing: decision-level or feature-level fusion?

2. Previous work

A number of approaches are based on multimodal affect sensing.

Kim and André [4] study affect sensing using fusion of physiological and acoustic data. The approach uses 77 features from the physiological modality, e.g. mean value, standard deviation, and ratio of max/min of physiological signals such as skin conductivity, electrocardiogramm and 61 features from the acoustic modality, e.g. mean, absolute extremum, root mean square, standard deviation of energy. It implements feature-level fusion, decision-level fusion and hybrid fusion. Using feature-level fusion, the approach computes a joint feature set from the two modalities for which a classifier is trained. In decision-level fusion, the output of separate classifiers for each modality is combined using a probabilistic approach. Hybrid fusion utilizes the output of the feature-level fusion as an additional input to the decision-level fusion. The best results were obtained by feature-level fusion in combination with feature selection.

Busso and colleagues [1] describe an approach to fusion of the facial expression modality and the acoustic modality at the feature level and the decision level. The examined corpus includes 258 emotional sentences annotated with 4 emotions. As acoustic features, they use features based on the mean, standard deviation, range, maximum, minimum and medians of pitch and energy. As visual features, the approach uses a 10-dimensional feature vector representing positions of particular 3D face markers. In feature-level fusion, the approach merges features from both modalities; in decision-level fusion, the approach utilizes either the best 10 features selected using SDFS or uses posterior probabilities and weights modalities. By fusing the facial and acoustic modalities the approach achieves an improvement of accuracy rates while the performance of the two fusion approaches was similar.

Schuller and colleagues [9] study fusion of acoustic and the lexical features. Experiments are performed on 1,144 phrases from seven American movie scripts. The approach uses 276 acoustic features, for example, pitch and energy; the lexical features are lemmatized unigrams (Bag-of-Words) without 93 stopwords (articles, names, etc.). The performance of the emotion recognition could be slightly improved by integration of acoustic and lexical features.

Truong and Raaijmakers [10] describe an approach to automatic recognition of spontaneous emotions that relies on the acoustic and the lexical modalities. It uses acoustic features (mean, standard deviation, max-min, the averaged slope of pitch and intensity) and lexical features (N-grams and the speech rate). The approach analyses positive/negative emotions and presents both uni-modal results and results of fusion at the feature level obtaining slight improvement after fusion.

In summary, the fusion of multiple modalities led to an increase of recognition results. However, we extend previous approaches in several respects. First of all, previous work focused mainly on calculating higher classification results without investigating thoroughly the influences of the single modalities. Then, they

analyzed only a limited number of linguistic features, i.e. lexical, while we use lexical as well as stylometric and deictic features. Furthermore, we also consider the context of turns in dialogues and provide a clear visualization of fusion results by means of a tree representation. Hereby, we introduce a measure to estimate the possible upper bound for multimodal fusion on the decision (classification) level — the maximal multimodality value.

3. Experimental setting

We perform experiments in this study using the Sensitive Artificial Listener (SAL) corpus [6] which contains audio-visual data of four users communicating with one of four psychologically different characters: optimistic and outgoing (Poppy), confrontational and argumentative (Spike), pragmatic and practical (Prudence), depressing and gloomy (Obadiah). The characters try to draw the user into their own emotional state, thus eliciting emotional speech. Dialog turns in SAL are transcribed manually and emotions in the turns are annotated by 3-4 annotators with the FEELTRACE software [2] which allows for continuous annotation of the emotion dimensions (valence and arousal). In total, SAL contains 27 dialogs (672 turns).

We mapped FEELTRACE values of turns onto 5 emotional classes assigning each turn the majority vote of the annotators at the end of the turn. Since majority calculation was not always possible due to the missing agreement between the annotators, we extracted only 574 turns from the original corpus that corresponds to 85% of the entire corpus. The extracted turns are 176 turns with low valence and high arousal, 103 turns with low valence and low arousal, 123 neutral turns with valence and arousal around zero, 24 turns with high valence and high arousal, 148 turns with high valence and low arousal. We considered a sixth "undefined" class of emotions that corresponds to turns whose affect is unclear: the annotators do not agree about the affect of the particular turn. However, we had to discard this class since the number of turns of this class was very big and the resulting distribution of affect classes would not reflect what SAL characters were intended to induce.

To avoid computational complexity, we perform our experiments in two stages distinguishing between 4 information streams of speech: lexical, stylometric, deictic information streams from the linguistic modality, and the acoustic modality. In the first stage, the datasets of a particular stream are composed and classified. In the second stage, 10 best datasets of every stream are left for further experiments. Furthermore, since successive turns in SAL dialogs are semantically connected, we compose feature sets not only for the current turn but also for the current turn plus n preceding turns (the context of the turn). Hereby, we restrict to $n=7$ which we empirically found to be a beneficial value.

In the lexical information stream, unigrams are used as features for composition of 29 lexical datasets. A lexical dataset considers the most frequent words in SAL frequency list and contains s/n features where $s=2,033$ is the length of the frequency list in words and n is the dataset number. Hence, the first dataset of the lexical modality ($n=1$) contains 2,033 features; the second dataset ($n=2$) consists of $1,019$ words; the third dataset ($n=3$) contains 679 unigrams, and so on.

We composed 31 stylometric datasets that contain possible combinations of the following feature groups [3], [5], [7]: letters, word lengths, digrams, standard deviation of word length, and sentence lengths in words. A stylometric dataset contains at least 1 feature and at most 730 features. The sentence lengths were represented as *frequency* vector.

For deixis we consider 63 datasets with combinations of the following feature groups: demonstratives as determiners, demonstratives as pronouns as well as time references, place references, forms of the third person (references to persons or subjects as *he* or *it*), and 526 stopwords from the WEKA toolkit [13]. A deictic dataset contains at least 1 feature and at most 530 features. The features were evaluated as frequency vector.

The acoustic feature set contains 1,316 features based on pitch, signal energy, MFCCs, the short-term frequency spectrum, and the harmonics-to-noise ratio that are extracted using the EmoVoice software [12].

We discretize the values of acoustic features. Discretization is a data mining method of feature evaluation that maps values in particular intervals onto interval names. For instance, a value of feature *pitch_mean* can be mapped onto an interval name as follows: values in interval $(-\infty, 108.5)$ are interpreted as the name *Interval1*, values in interval $(108.5, 165.2)$ are interpreted as the name *Interval2*, values in interval $(165.2, 221.5)$ are interpreted as the name *Interval3*, values in interval $(221.5, -\infty)$ are interpreted as the name *Interval4*. Hence, the sequence $(120.9, 105.1, 187.3, 275.1)$ is interpreted as the sequence (*Interval2*, *Interval1*, *Interal3*, *Interval4*).

We compile 4 datasets of acoustic features: a dataset representing the current turn; a dataset representing the current turn and 7 previous turns; a dataset with discrete values representing the current turn; a dataset with discrete values representing the current turn and 7 previous turns. The supervised discretization of acoustic datasets is performed using the Fayyad and Irani's discretization filter in the WEKA toolkit.

The classifier used throughout the experiments is SVM from the WEKA toolkit; the results are averaged over classes.

4. Decision-level fusion

To perform fusion at the decision level, we calculate the majority vote from the results of the single

information streams. If no majority can be established, we gradually leave out the information stream with the lowest recall value in uni-modal recognition, until either

majority voting is possible or only one information stream remains.
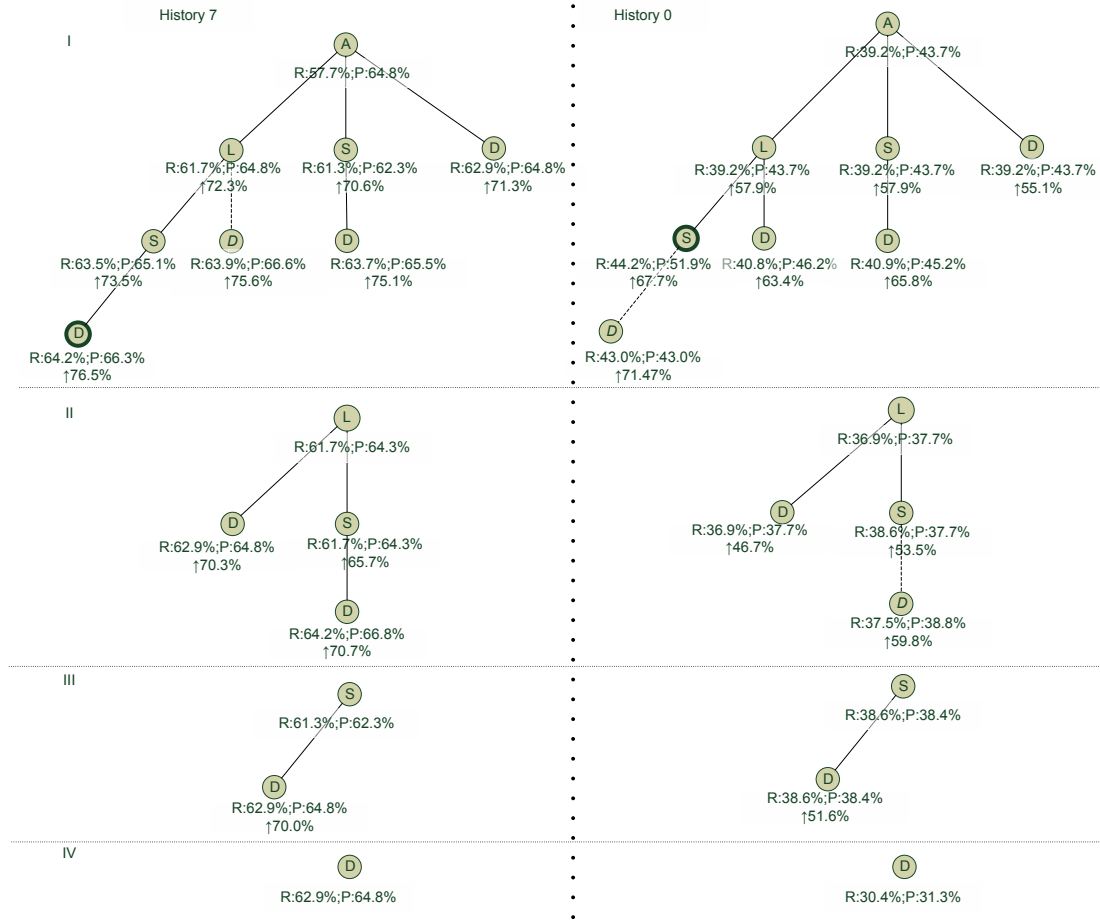

Figure 1: Decision-level fusion before discretization

In Figure 1, each path in the trees represents a dataset with the feature groups contained in this dataset starting with a different root node. For instance, the A-L-S-D path corresponds to a dataset with the acoustic (A), lexical (L), stylometric (S), and deictic (D) features. Each abstract I, II, III, IV corresponds to 4 datasets that are visualized using a different "root" information stream. The trees to the left show fusion results using the current turn in the context of the turns of history 7. The trees to the right show fusion results without history.

Results in the first row are denoted using the class-wise recall (R) and precision values (P). The second row denotes ↑<*maximal multimodality value*> where <*maximal multimodality value*> is the value that could be obtained in the case of perfect fusion when at least

one of the participating information streams would classify a particular instance correctly and the classification would rely on this information stream only. In other words, we understand the maximal multimodality value as an expectation of the maximal recall value that can be achieved using the participating datasets or as an anticipated upper bound of recognition (the coverage). Nodes that represent datasets with maximal recall values are shown in bold circles. Arcs that indicate decreasing recall values are dashed; the names of the corresponding nodes are italicized.

Result trees that correspond to values after discretization of the acoustic datasets are shown in Figure 2.
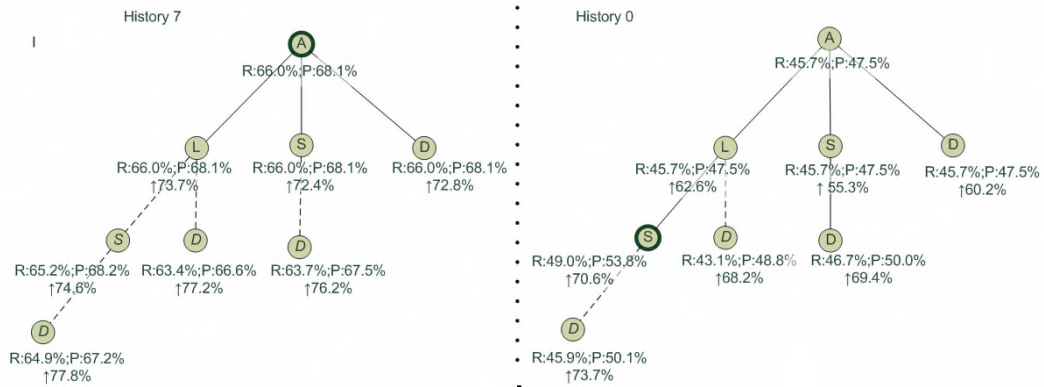
Figure 2: Decision-level fusion after discretization

Descriptions of nodes are the same as descriptions in Figure 1.

5. Feature-level fusion

Fusion at the feature level is performed by merging features of participating information streams into a single feature set (Figure 3).
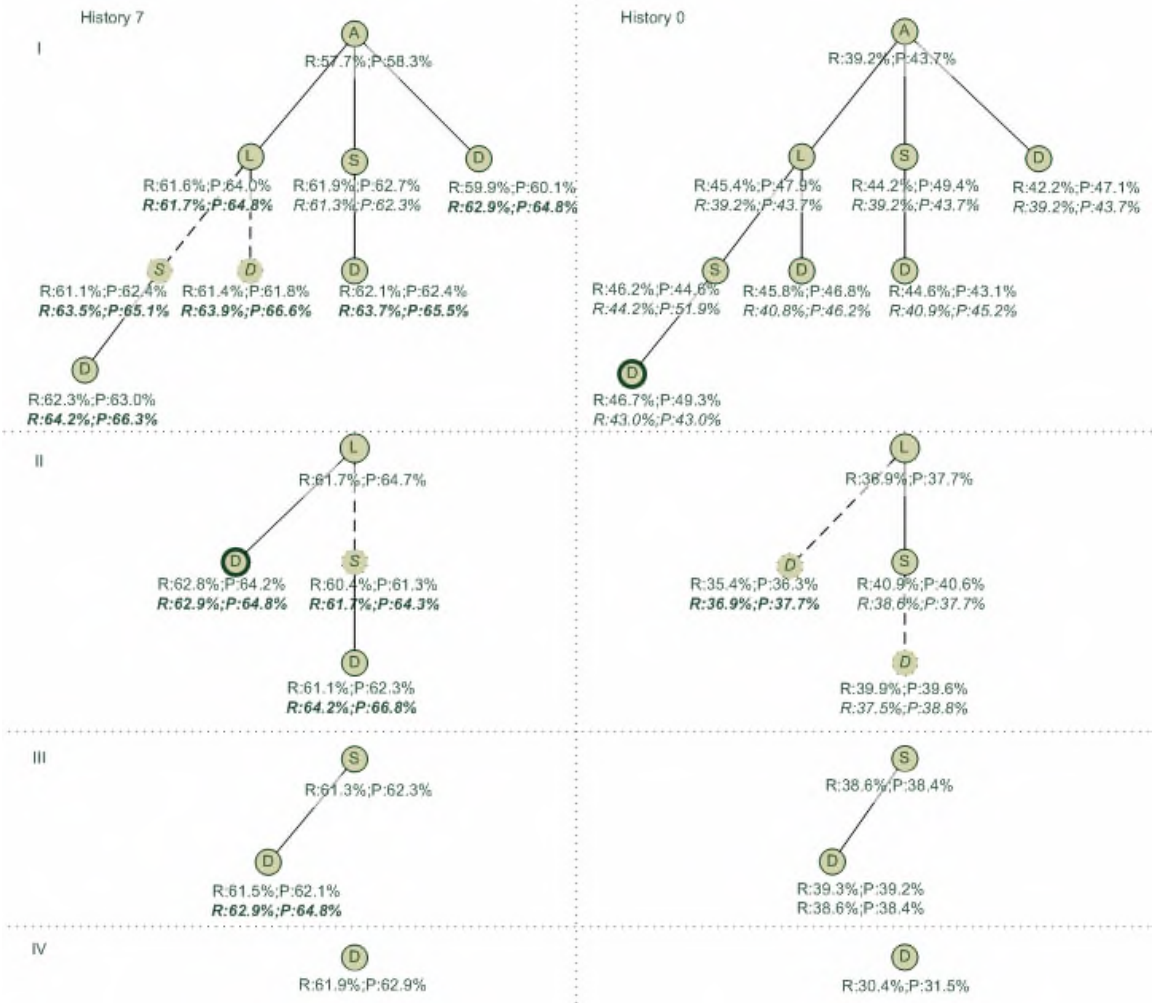

Figure 3: Feature-level fusion before discretization

Results are represented as trees similar to Figure 1 with a slight difference. The first row represents the recall value (R) and the precision value (P). However, the second row in italics shows the results (R, P) of the decision-level fusion once more in order to facilitate comparison of two fusion types. If the recall value in the feature-level fusion is greater than that in the decision-level fusion the second row is shown in bold. Note that some fusion results in Figure 3 indicate that the decision-level fusion was more beneficial for affect sensing than the feature-level fusion.

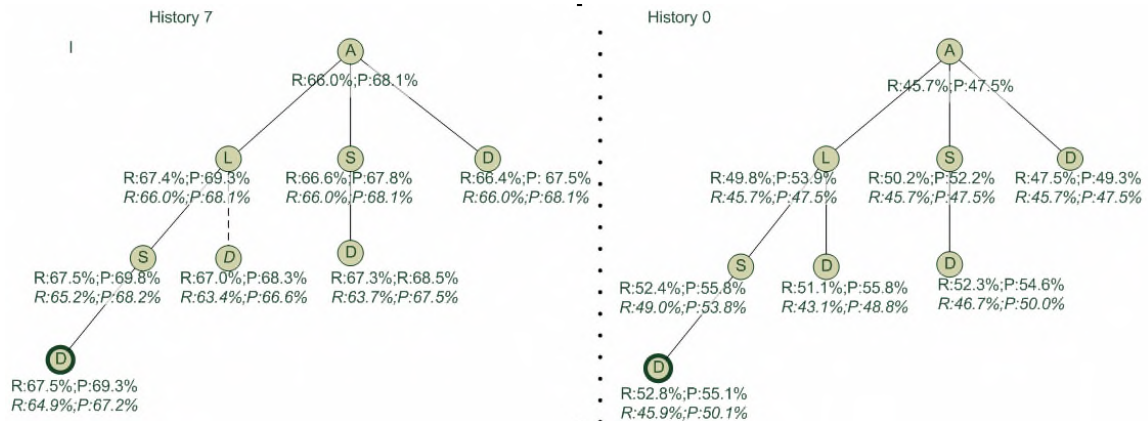Again, we also compiled trees that show results of fusion after discretization of the acoustic datasets (Figure 4).

Figure 4: Feature-level fusion after discretization

The descriptions of the trees are the same as the descriptions in Figure 3. Note that Figure 4 contains in contrast to Figure 3 no rows that denote that the decision-level fusion yielded higher results than the feature-level fusion.

6. Discussion

The fusion results confirm the following:

1. *Role of context:* Both for acoustic and linguistic features, we got better results when considering the context of a turn. However, this result may hold only to one corpus. Using a fixed number of turns might, however, be problematic in case where emotions rapidly change.

2. *Performance of the single information streams:* The discretization of acoustic features remarkably improves classification rates. After discretization, acoustic features score better than linguistic features. Without discretization, linguistic features score better than acoustic features if context is considered. The high impact of discretization is probably due to the size of the database which is relatively small. However, as we have shown on a bigger database [11], the effect still holds up to a considerable size.

3. *Fusion:* We were not in general able to improve recognition results by fusing linguistic and acoustic features. Only if no context was considered feature-level fusion led to better results than the analysis of the single information streams. However, when considering the context, the recognition rates obtained for the single information streams outperformed this result. Furthermore, we did not find neither fusion at the feature level nor at the decision level to be superior as both achieved about the same recognition results.

7. Conclusion

In this paper, we investigated decision-level and feature-level fusion and showed fusion results for a natural-language multimodal corpus. It turned out the discretized acoustic features outperformed linguistic features and that fusion of acoustic and lexical features was less beneficial than expected. In the paper, we introduced the maximal multimodality value as a means to assess the best possible results of decision-level fusion. In the analysis we conducted, this value goes to at most 77.8% demonstrating that a 100% recognition rate is not possible without improving the recognition rates of the single information streams. Therefore, we plan in the future to enhance results of the decision-level fusion by improving the initial information streams as well as adding new modalities in order to increase the coverage (the multimodality value), for instance, we plan to add the visual modality [8]. Also we plan to conduct research on weighting information streams and possibly also weighting feature sets in feature fusion. Furthermore, we will conduct feature selection to eliminate correlated or redundant features, and thus improve recognition results.

Acknowledgments

References

[1] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. Proceedings of the 6th international Conference on Multimodal interfaces. ICMI '04: 205-211. ACM, New York, NY. 2004.

[2]   Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. 'FEELTRACE': An instrument for recording perceived emotion in real time. Proceedings of the ISCA Workshop on Speech and Emotion: 19-24. Newcastle, Northern. Ireland. 2000.

[3]   Forsyth, R. S., & Holmes, D. I. (1996). Feature finding for text classification. Literary and Linguistic Computing, 11 (4): 163-174.

[4]   Kim, J., André, E. Emotion recognition using physiological and speech signal in short-term observation. Perception and Interactive Technologies. LNAI 4201: 53-64. Springer. 2006.

[5]   Kjell, B. Authorship determination using letter pair frequency features with neural network classifiers. Literary and Linguistic Computing, 9 (2): 119-124. 1994.

[6]   Kollias, S. ERMIS project. 2008.

[7]   Ramyaa, C., & Rasheed, K. Using Machine Learning Techniques for Stylometry. International Conference on Machine Learning; Models, Technologies and Applications. Las Vegas, USA. 2004.

[8]   Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., Schuller, B. 2008. *Towards responsive Sensitive Artificial Listeners.* Proceedings of Fourth International Workshop on Human-Computer Conversation. Bellagio, Italy.

[9]   Schuller, B., Müller, R., Lang, M., & Rigoll, G. (2005). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features Within Ensembles. Proceedings of Interspeech: 805-808. Lisbon, Portugal. 2005.

[10] Truong, K. P., Raaijmakers, S. Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features. Machine Learning for Multimodal Interaction: 161-172. Springer. 2008.

[11] Vogt, T., André, E. Exploring the benefits of discretization of acoustic features for speech emotion recognition. Proceedings of Interspeech. Brighton, UK. 2009.

[12] Vogt, T., André, E., Bee, N. EmoVoice — A Framework for Online Recognition of Emotions from Voice. In 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany,.LNCS 5078: 188-199. Springer. June 16-18, 2008.

[13] Witten, I. H., & Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann. 2005.