Classification of Emotional States in a Woz Scenario Exploiting Labeled and Unlabeled Bio-physiological Data

Martin Schels[1], Markus Kächele[1], David Hrabal[2], Steffen Walter[2], Harald C. Traue[2], and Friedhelm Schwenker[1]

[1] Institute of Neural Information Processing, University of Ulm, Germany
[2] Medical Psychology, University of Ulm, Germany
firstname.lastname@uni-ulm.de

Abstract. In this paper, a partially supervised machine learning approach is proposed for the recognition of emotional user states in HCI from bio-physiological data. To do so, an unsupervised learning preprocessing step is integrated into the training of a classifier. This makes it feasible to utilize unlabeled data or – as it is conducted in this study – data that is labeled in others than the considered categories. Thus, the data is transformed into a new representation and a standard classifier approach is subsequently applied. Experimental evidences that such an approach is beneficial in this particular setting is provided using classification experiments. Finally, the results are discussed and arguments when such an partially supervised approach is promising to yield robust and increased classification performances are given.

1 Introduction and Related Work

The reliability of a classifier heavily depends on the quality and quantity of the data, that was available for its construction. Unfortunately, in real world applications, it is often not trivial to design data bases where the data samples are exhaustively labeled. The main reason for this is that the general procedure of labeling data is often time consuming and expensive as it requires the knowledge of human experts.

There are several techniques in the literature, that aim at circumventing this issue by incorporating a machine-conducted labeling procedure: to make the annotation process more effectively, active learning is often used to guide a human expert during an annotation process. Hereby, the most informative sample from the unlabeled data, i.e. the one closest to a precomputed decision boundary, is selected by the algorithm and passed to an expert [4]. In order to conduct a fully automatic process, semi supervised learning can be applied: classifiers are directly used to annotate the unlabeled data. A classifier can label data for itself by choosing the most confident data samples and add them to the training set (self training) [15]. Another option is to use several classifiers in order to mutually select confident samples for the respective training data (co-training) [2,6,9].

F. Schwenker and E. Trentin (Eds.): PSL 2011, LNAI 7081, pp. 138–147, 2012.

In this contribution, we implement a further learning strategy to exploit unlabeled data in a classification process. The key idea is to infer the general structure of the application using methods of unsupervised learning [25]. This leads to a representation space using the cluster centers as reference system. For these computations all available data can be used. Such an approach appeared to be beneficial in previous work such as [19].

The remainder of this paper is organized as follows: The underlying data collection is described in Section 2 together with the employed features. Section 3 points to the general issues that occur in the application and introduces the proposed method in greater detail. The experiments and the respective results are shown in Section 4. Finally, in Section 5 these results are discussed and conclusions are drawn.

2 Data Collection

The data was collected in a Wizard-of-Oz study [7], which was conducted in order to investigate affective human computer interaction in the well established PAD space. The PAD model [17] defines a three dimensional annotation scheme of emotions using the three dimensions *pleasure*, *arousal* and *dominance*.

In this particular setting, the test persons were instructed to solve multiple games of concentration using a voice controlled interface. The successive games were used to induce different emotional states to the subject in the order sketched in Figure 1. To do so, different stimuli were presented to the subject deliberately: Different negative (dispraise, time pressure, wrongly or delayed executed commands, etc.) as well as positive (e.g. praise, easier game) behaviors of the computer interlocutor were presented. The subjects were passed through 5 sequences, which induce different states in the PAD space and the subjects each passed through these sequences twice in two successive sessions (see Figure 1 for details) [24]. Each of these sequences has a length of 3-5 minutes. Overall, 20 subjects (21 to 55 years, 10 male and 10 female) were passed through the experimental procedure twice and thus for every person two experimental sequences are available.

As a whole, 5 different channels were recorded at a sample rate of 512 Hz, namely blood volume pulse (i.e. heart rate), electromyography (attached to musculus zygomaticus and musculus currugator), skin-conductance and respiration. From these signals, various features were extracted on different time scales. Hereby, it is crucial to conduct a careful preprocessing procedure in order to remove artifacts but to retain the respective information. In general, a slow low- or band-pass filter is applied together with a linear piece-wise detrend[1] of the time series at a 10 s basis. In the following, a list of the extracted features per channel is provided. The preprocessing together with the time granularity is given in parentheses.

[1] i.e. subtracting piecewise a linear least-squares-fit from the respective chunk of the data.

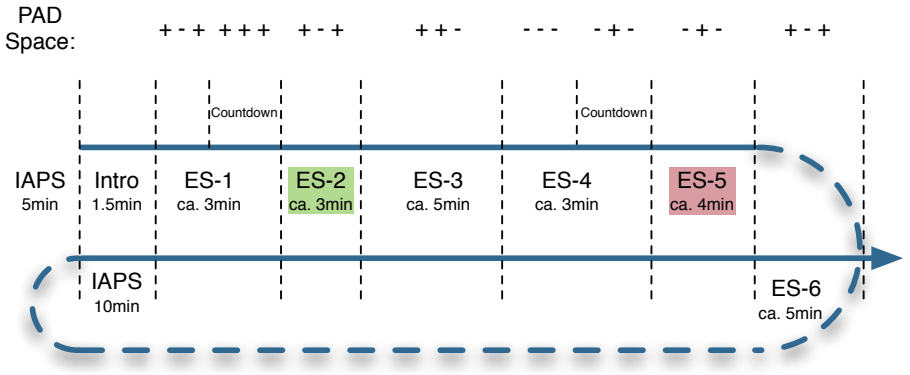

Fig. 1. Experimental design, including the expected position in the space. Experimental sequences ES-2 and ES-5, marked green and red respectively, are expected to induce the desired emotional states [24]. The top row in the figure indicates the intended label in PAD Space, whereby "+" signifies a high value in the respective dimension and "-" vice versa.

Blood volume pulse (BVP) is recorded from an optical sensor device, attached to a finger of the subject. The key to characterize the heart rate from the recorded blood volume pulse is to find the well known QRS complex in the signal e.g. as described in [16]. The following features are extracted (low pass filtered at 5 Hz, 25 s time window each) : Standard deviation of heart rate variability [18], standard deviation of RR-intervals [23], pNN50[2] [12], approximate entropy [13], RMSSD[3] and recurrence rate, Poincaré plot[4] [10] and power spectral density [26] of the signal.

The subject's respiration [3] is measured using a belt, that is wrapped around its breast and has a tension measurement device attached to. From this signal the following features are computed (low pass filtered at 0.15 Hz): Mean and standard deviation of the first derivatives (10 s time window), breathing volume, mean and standard deviation of breath intervals, Poincaré plot[4] (30 s time window each).

To record the electromyogram (EMG), 2 electrodes are attached to the skin near to the respective muscle. Thus electrical potential differences of about 500 μV are recorded. Hereby lies the information of contraction or relaxation of the muscle in the oscillation of the EMG signal. The following features were computed (bandpass filtered at 20 - 120 Hz, piecewise linear detrend): Mean of first and second derivatives (5 s time window), power spectrum density estimation [26] (15 s time window).

[2] The pNN50-measure equals the proportion of occurrences of changes in RR-interval duration of two consecutive RR-intervals that differ more than 50ms.
[3] Square root of the mean of squared successive differences of RR-intervals.
[4] Ratio of the axes of the fitted ellipse.

Skin conductance is measured (SCL) using 2 electrodes, where constant electrical current of 10 μA conducted. The respective resistance is then determined by the sweat, a subject oozes. The following features are extracted for this signal (low pass filtered at 0.2 Hz): mean and standard deviation of first and second derivative (5 s time window), mean peak occurrences, average peak-height (20 s time window each) [5].

In the following, the task to solve in the context of this paper is to discriminate the samples of ES-2 and ES-5. These two experimental sequences are designed to elicit rather complementary emotions: "high pleasure/low arousal/high dominance" versus "low pleasure/high arousal/low dominance" (compare Figure 1, top row) – or short positive vs. negative emotion. The according stimuli that are presented to the subject were praises and a small board of concentration and hence an easy play for the positive sequence. In case of the negative class, the user is given a bigger board and only displeasing feedback is given: e.g. the user is criticized for his execution of the game and the subject is exposed to time pressure.

3 Problem Statement and Proposed Method

An application as described above arises several severe issues from a machine learning perspective. Based on the design of the psychological experiment, the overall samples that are labeled accordingly are very rare. When attempting to compute reasonable features from the given data, the respective time window has to be chosen over several seconds. Due to the high differences of physiology over different subjects, the given application encourages the commands for a personalized setting in the training of classifiers. This further toughens the lack of data.

Further, when evaluating such kind of data it is not recommended to use some kind of "leave one sample out" technique to evaluate a statistical model. The employed sensors show a distinct characteristic over time and as the labels are heavily correlated by definition to time, this would imply a severe bias in the results. This implies in our application that it is necessary to train and test using data from different sessions. Hence, it is highly desirable to make use of all available data from all experimental sequences recorded from a subject. Unfortunately this data is not labeled in the respective classes (compare Figure 1). On the other hand, it is still data from the same domain. The goal is now to incorporate all available data into the construction of a classifier for the considered two classes.

To do so, it is rather intuitive to refer to techniques of unsupervised learning. The key idea is now to neglect the actual class labels for the samples and to process all available data using a unsupervised technique - such as k-means or Gaussian mixture models. In order to solve the actual classification problem a further learning step is implemented: Based on the computed partitioning of the data, an "activation value" of the cluster centers for the data samples is computed. This activation could either be computed by a distance measure with

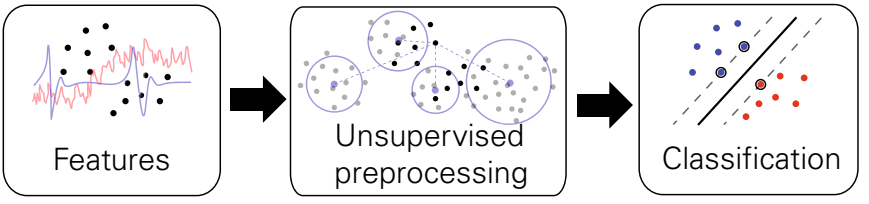

Fig. 2. In order to incorporate the unlabeled data into the classification, an unsupervised learning step is implemented. Thus, the data is transformed into a new representation, in which the actual classification is conducted.

respect to a cluster center in case of a partitioning algorithm is used, or the posterior probability of a mixture component of a fitted generative model. This results in a new representation of the data of the same dimensionality as number of cluster centers. Based on this new feature vector, a classification on the initial label is conducted using standard supervised machine learning approaches. This procedure is sketched in Algorithm 1.

Algorithm 1: Proposed algorithm in pseudo code.

Input:
- Labeled data $L = (l_i)_{i=1...M}$
- Respective labels $Y = (y_i)_{i=1...M}$
- Unlabeled data $U = (u_j)_{j=1...O}$
- Number of cluster centers k,

compute k local densities or prototypes $p_1, \ldots, p_k$ using $L \cup U$;
foreach $l_i \in L$ **do**
  $l'_i = G_{p_1,\ldots,p_k}(l_i) \in \mathbb{R}^N$;
  G is a distance or similarity measure, and N is a natural number depending on the specific structure of G
  examples:

  (a) $N = k$, and $G_{p_1,\ldots,p_k}(l) = \mid p_i - l \mid$
  (b) $N = k$, and $G_{p_1,\ldots,p_k}(l) = (\exp(- \mid p_i - l \mid / \sigma_i)$
  (c) $N = k(k-1)/2$ and $G_{p_1,\ldots,p_k}(l_i) = \min_{i,j}(\mid p_i - l \mid, \mid p_j - l \mid)$

end
Train classifier F on $((l'_i)_{i=1...M}, Y)$;
Output: F

To classify an unseen data sample, it has to be transformed into the new representation. This is done analogously to the training procedure by calculating the activation score: These values are computed with respect to the computed local density or the respective prototype and the obtained new representation is classified.

4 Experiments and Results

In this section the conducted experiments and the obtained results are described. An important step in our approach is the choice of the unsupervised learner: We decided to evaluate the well known k-means algorithm, neural gas [11] and Gaussian mixture models (GMM) trained through the well known expectation maximization (EM) algorithm. In case of neural gas and k-means, the number of cluster centers is chosen to be 15% of the size of the training set. Due to numerical issues the number of mixture components for the GMM is set constantly to 4 and a regularization constant of 0.001 was fixed. Generally, the euclidean distances to the cluster centers were used as new representation except for GMM, where the posteriori probability for every Gaussian mixture component was computed.

For the supervised part of the proposed architecture, a support vector machine (SVM) learning approach was used [1]. To be precise, in this work ν-SVM as described in [20] was used using an RBF kernel function. To compare the experimental results, we also conducted a purely supervised reference experiment, where the training of the classifier is conducted only on the accordingly labeled data. For this experiment, we used the ν-SVM approach with an RFB kernel as well.

In Section 3 the general issue of testing a classifier in this application appropriately was mentioned briefly. To circumvent this issue and in order to ensure proper results, the partitioning in training and testing data are separated by session per class. But to increase the possible settings for testing, the whole experimental sequences are permuted in all possible combinations. This leads to four settings of testing and training sets per subject.

The features described in Section 2 are extracted not only from different modalities but also in different time scales. Hence, the classification study was conducted in six different experiments, grouping the data by feature and size of the time window: For the EMG, features that govern in time domain (derivatives of the signal and related) are grouped together as well as features obtained from the power spectrum. Also for the skin conductance two groups of features were defined for classification: The statistics over the derivations are processed in a different classifier than the statistics of the peaks of the signals. In case of BVP and respiration such a partitioning is not necessary as the time windows of all extracted features are the same.

The performances of the classifiers are reported in Table 1. As the distribution of classes in the data is imbalanced (compare 1) not only the accuracy are reported, but also the F1 scores for ES-2. The numbers in the rows are mean values over all subjects and every classifier is evaluated 80 times each. Generally, the numbers are relatively low, which is not surprising as the application is rather challenging together with the general lack of data. It can be observed that the two classifiers using EMG features perform best with an accuracy up to 0.53. Also the classification on respiration features performs well (0.51 accuracy). All classifiers avoid to produce one-sided classification result, i.e. it does not constantly decide for the class having the higher a priori probability, which is indicated by the F1 scores.

Table 1. Accuracies and F1 scores for ES-2 for the 6 classifier configurations averaged over all subjects and 80 trials per subjects. The row-wise maximal values for both values are highlighted in bold font. The last row shows the averages over all classification trials.

Feature combination	GMM		Neural-Gas		K-Means		purely supervised	
	acc.	F1	acc.	F1	acc.	F1	acc.	F1
EMG (derivatives)	0.529	0.404	**0.530**	**0.428**	0.486	0.389	0.502	0.394
EMG (power spectrum)	**0.531**	0.404	0.514	**0.430**	0.461	0.366	0.458	0.342
SCL (derivatives)	0.431	**0.325**	**0.431**	0.300	0.415	0.321	0.424	0.323
SCL (inter-peak statistics)	0.437	0.356	0.448	0.399	**0.451**	**0.399**	0.421	0.355
BVP	0.475	0.363	0.437	0.355	0.455	0.366	**0.483**	**0.392**
Respiration	0.449	0.347	**0.510**	**0.447**	0.484	0.407	0.503	0.368
Average	0.475	0.366	**0.479**	**0.393**	0.459	0.374	0.465	0.362

Table 1 provides some arguments, that the unsupervised preprocessing does provide benefits for the classification: In 3 of 6 cases of the classifiers, the partly supervised method using neural gas outperforms the others. Further, comparing all partly supervised experiments to the purely supervised case, it performs best in 5 of 6 cases on average. Also, when averaging over all test runs, there is a slight preference for the clustering preprocessing approach using neural gas but also using GMM.

A ranking-like experiment is conducted, where it is counted how often a classifier outperforms all others for every individual subject averaged over all 80 trials. The results of this are reported in Table 2 as fractions of all comparisons. This consideration reveals a slight advantage of the GMM based partially supervised classifier: It outperforms the others in 32% of the cases. Especially for the features from EMG, which performed best in Table 1, such an approach appears to be beneficial.

Table 2. For every classifier it is shown how often it outperforms all others. The lines of the table show different feature combinations.

Feature combination	GMM	Neural-Gas	K-Means	purely supervised
EMG (derivatives)	**41.1**%	17.7%	17.7%	23.5%
EMG (power spectrum)	**47.1**%	23.5%	5.9%	23.5%
SCL (derivatives)	23.5%	17.7%	23.5%	**35.3**%
SCL (inter-peak statistics)	29.4%	17.7%	**35.3**%	17.7%
BVP	**35.3**%	5.9%	23.5%	**35.3**%
Respiration	17.7%	**35.3**%	11.8%	**35.3**%
Average	**32.4**%	19.6%	19.6%	28.4%

5 Discussion and Future Work

In this work a partially supervised machine learning approach has been proposed and applied to the classification of bio-physiological time series. In this application, only few data is available in the considered classes, but there is differently

annotated data at hand, that did arise in the overall recording process. The goal was to incorporate these samples into the classification process. To do so, we propose to use an unsupervised learning approach as a preprocessing step. Three different learning strategies have been evaluated in this context: k-means, neural gas as clustering approaches and GMM to estimate the probability distribution. Thus the data was transformed in a new representation using the activation per prototype or mixture component. Using the partitioning algorithm, the euclidean distance has been used, while for the GMM the posterior probability per mixture component is used. The experimental results reveal a slight advantage over the purely supervised reference method of such an approach in this application.

In order to provide a rationale of why the proposed method works, the reader is pointed to the the well known RBF networks. There exists a big research community exploring how to improve the training of a network from given data by finding a proper initialization [8,14,21,22]. Hereby, the aim is to make the results more stable and also to speed up training. A typical approach is to pretrain the hidden RBF-layer in an unsupervised fashion by clustering or vector quantization. Afterwards, the network is finally trained by either solely creating a perceptron for the output layer or back propagation for the whole network. The unsupervised step in our approach can be regarded as some sort of initialization of a "hidden layer" using all available data. Thus, the distributions of data can be estimated more reliably. After that, a second "output layer" is created with only the labeled data at hand.

Adding additional data the way we did in our experiments, i.e. data, that is not from the same categories is of course only promising under certain conditions. If the samples of data resolved into clearly delimited classes, where the probability density functions for the different categories are non-overlapping, adding data from a very different partition would hardly be reasonable. But in many real world applications, this optimal setting for a classifier is not really present: Often the data decomposes into severely overlapping distributions. There are also applications, where the particular classes are not (yet) irrevocably defined or such a definition is simply not possible due to distinct properties. Both circumstances are at hand in the application described earlier: On the one hand, the features that can be extracted from the bio-signals can be considered relatively weak compared to the intended – quite ambitious – objective. On the other hand, even though the induction of the intended emotion succeeds in the average, it is not guaranteed by any means that every particular sample is correctly labeled.

The relatively small accuracies reported in Table 1 could be regarded as a major flaw of this contribution. There are two major ways to heal this issue: There are still 6 individual classifiers that are evaluated in this study. These classifiers should be further combined in order to enhance a frame wise classification process. This has to include of course some kind of alignment of the different time windows that are used in order to get a coherent classification. Another promising approach is to integrate the decisions of the classifiers over

time [24]. On the other hand, one will then have to solve a general segmentation problem in order to discriminate the sequences.

Further, calculating the new representation of the data samples creates the opportunity to define mappings into higher dimensional spaces. This could, for example, conducted as sketched in Algorithm 1 at example (c), where pair-wise comparisons are used to build the new representation. Thus, it might be more likely to find a proper linear separation of the respective classes.

Acknowledgments. This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems", funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl-Zeiss Foundation.

References

1. Bennett, K.P., Campbell, C.: Support vector machines: hype or hallelujah? SIGKDD Explor. Newsl. 2, 1–13 (2000)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100 (1998)
3. Boiten, F.A., Frijda, N.H., Wientjes, C.J.: Emotions and respiratory patterns: review and critical analysis. International Journal of Psychophysiology 17(2), 103–128 (1994)
4. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. J. Artif. Int. Res. 4, 129–145 (1996)
5. Darrow, C.W.: The equation of the galvanic skin reflex curve: I. the dynamics of reaction in relation to excitation-background. The Journal of General Psychology 16(2), 285–309 (1937)
6. Hady, M.F.A., Schwenker, F., Palm, G.: Semi-supervised learning for tree-structured ensembles of RBF networks with co-training. Neural Networks 23(4), 497–509 (2010)
7. Kelley, J.F.: An empirical methodology for writing user-friendly natural language computer applications. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 193–196 (1983)
8. Kestler, H.A., Schwenker, F., Hoher, M., Palm, G.: Adaptive class-specific partitioning as a means of initializing RBF-networks. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 46–49 (1995)
9. Ling, C.X., Du, J., Zhou, Z.H.: When does co-training work in real data? In: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 596–603 (2009)
10. Marciano, F., Migaux, M., Acanfora, D., Furgi, G., Rengo, F.: Quantification of Poincaré maps for the evaluation of heart rate variability. Computers in Cardiology, 577–580 (1994)
11. Martinetz, T., Schulten, K.: A "Neural-Gas" Network Learns Topologies. Artificial Neural Networks I, 397–402 (1991)
12. Mietus, J.E., Peng, C.K., Goldsmith, R.L., Goldberger, A.L.: The pNNx files: reexamining a widely used heart rate variability measure. Heart 8, 378–380 (2007)

13. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences of the United States of America 88(6), 2297–2301 (1991)
14. Ros, F., Pintore, M., Deman, A., Chrtien, J.R.: Automatical initialization of RBF neural networks. Chemometrics and Intelligent Laboratory Systems 87(1), 26–32 (2007)
15. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: Proceedings of the Seventh IEEE Workshops on Application of Computer Vision, pp. 29–36 (2005)
16. Rudnicki, M., Strumiłło, P.: A Real-Time Adaptive Wavelet Transform-Based QRS Complex Detector. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4432, pp. 281–289. Springer, Heidelberg (2007)
17. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
18. Sayers, B.: Analysis of heart rate variability. Ergonomics 16(1), 17–32 (1973)
19. Schels, M., Schillinger, P., Schwenker, F.: Training of multiple classifier systems utilizing partially labeled sequences. In: Proceedings of the 19th European Symposium on Artificial Neural Networks, pp. 71–76 (2011)
20. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Comput. 12, 1207–1245 (2000)
21. Schwenker, F., Kestler, H., Palm, G., Hoher, M.: Similarities of LVQ and RBF learning-a survey of learning rules and the application to the classification of signals from high-resolution electrocardiography. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 646–651 (1994)
22. Schwenker, F., Kestler, H.A., Palm, G.: Three learning phases for radial-basis-function networks. Neural Netw. 14, 439–458 (2001)
23. Simson, M.: Use of signals in the terminal QRS complex to identify patients with ventricular tachycardia after myocardial infarction. Circulation 64(2), 235–242 (1981)
24. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal Emotion Classification in Naturalistic user Behavior. In: Jacko, J.A. (ed.) HCI International 2011, Part III. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)
25. Webb, A.R.: Statistical Pattern Recognition. John Wiley and Sons Ltd. (2002)
26. Welch, P.: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics 15(2), 70–73 (1967)