March 24, 2017

Post-Processing of Machine Classifier Output for Object Classification

Eric Nichols

Sourish Chaudhuri

Post-Processing of Machine Classifier Output for Object Classification

ABSTRACT

Machine classifiers are typically trained using labeled data sets. If the training data set has categories of objects that naturally co-occur, the machine classifier may have difficulty in distinguishing those categories. For example, audio streams often contain instances of sounds that occur simultaneously; e.g., speech and laughter. In this example, the different sounds are the objects that are to be classified. A machine classifier trained with such audio streams generates false positives; e.g., conflates speech with laughter, if the training data set does not label speech separately from laughter. The difficulty of obtaining well-labeled training sets compounds the problem of misclassification. For example, most transcriptions of audio streams containing laughter also include speech in close proximity, since laughter occurs just after speech; e.g., at the end of a joke. Furthermore, humans that produce training data typically annotate rather long audio segments at once, without specifying precise times for each word or audio event, so segments that contain laughter typically include both "speech" and "laughter" without labeling exactly when each occurred. This disclosure describes techniques to improve classification accuracy that are applicable for machine classifiers that act on any type of data; e.g., video, documents, images, etc.

KEYWORDS

- Machine classifier
- Training data
- Event detection
- Co-occurrence

BACKGROUND

Viewing online video is a popular activity. When captioned, online videos are accessible to a larger audience. Generating captions for large quantities of videos that are uploaded daily is a formidable task, often accomplished with the use of automated captioning systems. Audio streams (e.g., from online videos) contain a variety of sounds; e.g., speech, laughter, applause, whistles, etc. The different sounds in the audio stream are the objects that are to be classified. For a machine classifier to accurately transcribe audio (e.g., for it to identify laughter in the audio stream and transcribe it as such in an automatically generated caption) it is important that the classifier distinguish accurately between various categories of sounds.

A machine classifier sometimes generates false positives; e.g., it confuses one category of sound for another. For example, a classifier may report "laughter" in a segment that contains only speech, or report both "laughter" and "speech," when only speech is present. Such confounding of categories occurs because the data set used for training the classifier often has imprecise or weak labels for various categories of sounds. To some extent, imprecision in the labeling of training data is unavoidable. For example, if an audio segment contains both speech and laughter, a human transcriber or labeler is likely to transcribe the speech and include the word "laughter" in the transcript of the segment. Since transcribed segments have a somewhat long duration (e.g., 2-3 seconds) relative to the individual transcribed events, segments that contain laughter often end up labeled as both "speech" and "laughter" without any distinction about the relative order or overlap between the two categories within the segment. Further, such transcription does not specify the time intervals that contained pure speech, pure laughter, or both. Training segments containing pure laughter are therefore uncommon.

DESCRIPTION

This disclosure describes techniques to separate categories of objects that co-occur naturally in an audio stream; e.g., speech and laughter. A machine classifier that operates on a time-varying input stream — for example, an audio or video stream — produces a vector of values at every time step. Each value represents a determined probability that a certain category of object is present in the processed portion of the stream. For example, a machine classifier operating on an audio stream produces an *N*-dimensional vector corresponding to sound categories such as “applause,” “whistle,” “speech,” “laughter,” “sigh,” “music,” “ring,” “buzz,” etc. For example, when the *N*-dimensional vector is [0.01, 0.00, 0.50, 0.49, 0.00, 0.00, 0.00, 0.00, 0.00], then the classifier estimates that the probability of applause being present is 1%, the probability of a whistling sound is 0%, the probability of speech is 50%, the probability of laughter is 49%, and the probability of all other categories of sound is 0%. While the probabilities in this example add up to 100%, it is also possible to independently generate the probability for each category. When the probabilities are independently generated, the sum of probabilities does not necessarily add up to 100%.

At a successive time step (e.g., 10 milliseconds later) another vector is generated that contains probabilities corresponding to sound categories for the next portion of audio. Due to the naturally high co-occurrence in training data of certain categories of objects (e.g., sound and laughter) a classifier under test conditions often reports high probabilities for both sound and laughter, even if the audio stream includes just one of the two categories.

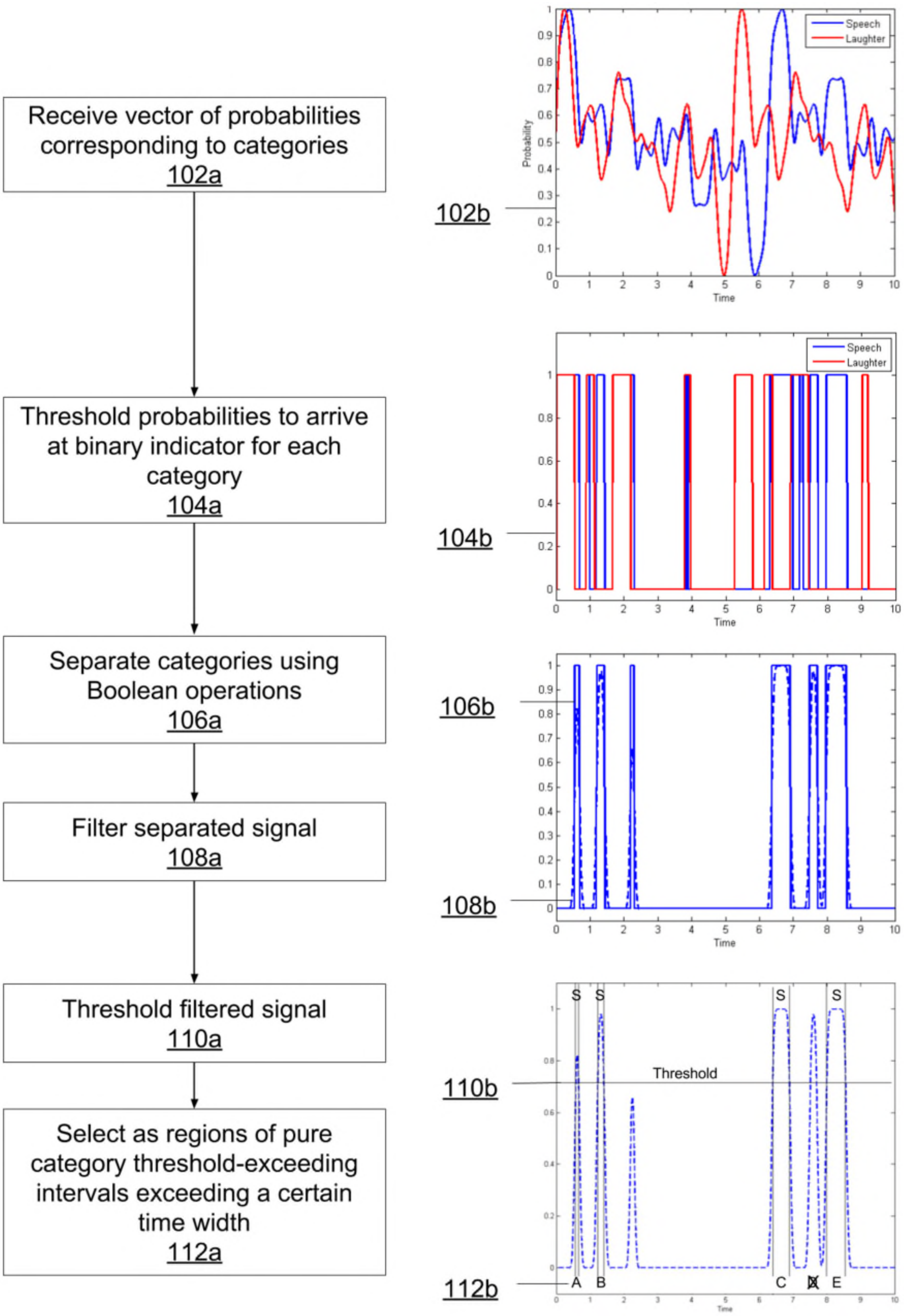
Receive vector of probabilities corresponding to categories
102a
Threshold probabilities to arrive at binary indicator for each category
104a
Separate categories using Boolean operations
106a
Filter separated signal
108a
Threshold filtered signal
110a
Select as regions of pure category threshold-exceeding intervals exceeding a certain time width
112a
102b
104b
106b
108b
110b
112b
Speech
Laughter
Probability
Time
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
4
5
6
7
8
9
10
Threshold
S
S
S
S
A
B
C
~~D~~
E

Fig. 1: Separating categories that confound a machine classifier

Fig. 1 illustrates an example process (and corresponding signals) to separate categories of objects (e.g., speech and laughter) that are conflated by a machine classifier. A vector of probabilities corresponding to each category is received at each time point (102a). Values of two elements of this vector — "speech" (in blue) and "laughter" (in red) — are illustrated (102b) against time. Other elements of the vector (e.g., "applause", "whistle", etc.) are present but omitted for the purpose of clarity. It is seen from the example of 102b that there are several time intervals when the speech and laughter probabilities are nearly equally high, illustrating the problem of conflating categories.

Probability signals for each category are thresholded (104a) to obtain two-level signals (104b) that indicate the presence or absence of a category. For example, if the threshold for laughter is 0.6, then time intervals with laughter probability greater than or equal to 0.6 are set to one (as shown in 104b), and time intervals with laughter probability less than 0.6 are set to zero. Thresholds for each category can be set independent of other categories. The two categories are separated (106a) using Boolean operations on the two-level signals to obtain a two-level presence/absence indicator signal (106b) for a single category. A Boolean operation for separated speech is, for example, *separated-speech* = (*speech*) AND (NOT (*laughter*)), where *separated-speech* represents the signal (106b) containing segments of pure speech and no laughter. In the above equation, *speech* and *laughter* represent respectively the blue and red thresholded signals (illustrated in 104b). Although the signal for separated laughter is not shown, a Boolean operation for separated laughter is, for example, *separated-laughter* = (*laughter*) AND (NOT (*speech*)).

Similarly, Boolean operations that separate any desired category, e.g., “applause”, “crying”, “whistle”, etc. are defined. A general Boolean expression for a separated category is, for example, as follows:

separated-desired-category =
(*desired-category*) AND NOT (*category-1* OR *category-2* OR *category-3* OR … *category-N*).

In the above equation, a *desired-category* is a particular category such as, for example, “laughter”, and *category-1* through *category-N* are other categories that appear in the audio stream and are confounded with the desired category, such as “crying”, “speech”, etc. Other Boolean expressions — e.g., that select up to 2 of *N*, up to 3 of *N*, etc. categories — can also be used.

After the separation of categories, the separated speech signal (106b) is filtered (108a) to obtain a filtered separated signal (108b, shown as dashed-blue). The filtering converts the binary signal to a smoothened signal. The filtered signal is thresholded (110a) using a threshold (110b) such that only time-intervals above threshold are deemed to contain pure category. Thus, time-intervals A, B, C, D and E, during which the filtered signal exceeds threshold, are deemed to contain, for example, pure speech. Further (112a), intervals that are of insufficient width, for example, interval D, are removed. Thus intervals that contain the pure separated category (112b) are deemed to be A, B, C and E. Additionally, time-intervals that occur close to each other and that contain a separated category are concatenated.

An alternative approach to separate categories is to subtract the probabilities of two categories. For example, a quantity $laughter'$ is defined as follows:

$$laughter' = max(0, laughter - speech),$$

where the magnitude of $laughter'$ is an indicator of pure laughter, and $laughter$ and $speech$ are respectively raw probabilities for the presence of laughter and speech, as generated by the machine classifier. The $max()$ operation is used to restrict $laughter'$ to a positive value. However, this approach may not be suitable, e.g., when the two categories have different prior probabilities.

Another approach to separate categories accurately is to train the machine classifier with clean and strongly-labeled training data. For some applications, substantial manual effort will be required to generate such data, which makes this approach expensive and time consuming. Rather, techniques of this disclosure can be used to automatically separate categories and thereby generate new training data that bears relatively strong labels. Training data thus generated can be sent to human labelers to develop cleaner training data.

While the examples described above refer to audio segments, the techniques described are applicable for any type of data in which multiple object categories are identified with the use of machine classifiers. For example, such data can include video or still images, documents, etc.

CONCLUSION

Machine classifiers are often unable to accurately and automatically separate categories of objects that naturally co-occur; e.g., speech and laughter in an audio stream. This is often due to insufficient diversity or bias in training data that is used to train machine classifiers. Techniques disclosed herein apply thresholding, Boolean operations, and filtering on the output of a machine classifier to separate categories of objects that confound the classifier. The techniques are simple to implement, require no changes to the machine classifier, reduce false positives, and improve object classification precision.