Post-Hoc Interactive Analytics of Errors in the Context of a Person Discovery Task

Pierrick Bruneau*, Mickaël Stefas*, Johann Poignant†, Hervé Bredin† and Claude Barras†
*LIST, L-4362 Esch-sur-Alzette
†LIMSI - CNRS, F-91405 Orsay

***Abstract*—Part of the research effort in automatic person discovery in multimedia content consists in analyzing the errors made by algorithms. However exploring the space of models relating algorithmic errors in person discovery to intrinsic properties of associated shots (e.g. person facing the camera) - coined as post-hoc analysis in this paper - requires data curation and statistical model tuning, which can be cumbersome. In this paper we present a visual and interactive tool that facilitates this exploration. A case study is conducted with multimedia researchers to validate the tool. Real data obtained from the MediaEval person discovery task was used for this experiment. Our approach yielded novel insight that was completely unsuspected previously.**

I. Introduction

This paper addresses a recurrent problem in person identification and discovery tasks such as MediaEval [1]: the post-hoc analysis of the errors made by identification algorithms, i.e. relating algorithmic errors to *intrinsic properties* of the associated media segments (e.g. acoustic or video properties). The goal is to explain why a given algorithm has succeeded or failed in retrieving a given shot as expected. In other words, are there acoustic or visual shot properties explaining for the *retrieval state* of a shot w.r.t. a queried speaker (success or error/failure), and which are they?

Being able to answer the question above may help task managers identify shot characteristics that are key to error patterns by systems participating to a competitive task, and give hints to better balance the difficulty of future issues of the task. Performing this analysis on two or more runs simultaneously may also support participants in verifying if variants of their processing pipeline behave as expected.

The effectiveness of shot properties (equivalently referred to as *features* in this work) to explain a retrieval state can be estimated in the context of predictive modelling: we consider that properties used as input by models predicting the retrieval state of shots (commonly referred to as *target* in the supervised learning terminology) with high accuracy can reveal strengths, weaknesses, and more generally sensitivity of retrieval algorithms.

The main contributions in this paper are:

- original statistical building blocks that support multiple targets and classification algorithms, towards the stepwise selection of features with high influence on the targets,
- linked target and feature views that adapt prior work in information visualization to the applicative context at hand and the above-mentioned statistical background.

In the MediaEval person discovery task, participant runs were converted to a database of couples *(shot ID, name)*, each associated with a distinct retrieval state (success of failure). Section IV discloses a case study where real users loaded this data in our tool.

II. Related Work

A. Model Selection

Models such as decision trees, already used in a post-hoc analysis [1], incorporate greedy selection and pruning techniques [2]. In [1] an oracle resulted from the combination of retrieval results of all participant runs for a subset of 2204 elements in the database. The retrieval state of the oracle is positive if at least one participant run retrieved the respective couple successfully. Cross-validation was used to estimate the Gini importance of properties [2], that indicates the tendency of decision tree models to retain a given property, along with its discriminative power.

B. Interactive Feature Selection and Ranking

Visual interaction can facilitate interpretation of ranks and scores returned by multiple feature selection algorithms. [3] considers this task in the context of multiple classification algorithms with cross-validation, applied to a data set with a single target. They introduce a novel pie-chart like glyph with inward-growing bars, which aggregates the ranks of several feature selection indexes and cross-validation folds, with respect to a large number of features. This enabled domain experts to quickly compare feature selection algorithms, and combine selections made by several of them. However, the considered scoring methods (e.g. information gain, Fisher score, odds ratio) estimate the predictive ability of a property independently of a given model and properties it already uses: they do not account for combined linear (i.e. correlations) and non-linear effects of properties in predictive models.

In section II-A, the Gini importance has been computed as a result of a decision tree learning algorithm. This score accounts for property interactions, but cannot be generalized to models beyond decision trees. The approach in [4] relies on ranked lists to effectively convey associated feature scores. They facilitate the combination of multiple scores in a single value using stacked bar charts with feature types mapped to categorical colors.

C. Stepwise Feature Selection

[5] uses visualization and interaction to perform stepwise feature selection for linear regression. This approach starts from a trivial model (i.e. only the intercept), and features can be added one at a time, with consequences in terms of model quality (R^2) shown interactively. The approach does not implement a feature selection method *per se*: it is up to user trial and error, i.e. inclusion of a feature leading to minor R^2 improvement indicates there is little interest in retaining it. Also they do not concern themselves with cross-validation folds to establish more robust estimates.

The likelihood ratio test can also be used to assess the statistical value of including a feature in a model [6]. The advantage of this test over the approach in [5] is that it can be used for any regression or classification model provided that the latter can be expressed as a probabilistic loss function. In that sense it is more general than the R^2, that is bound to the linear regression model. Each feature is assigned with a score that has a probabilistic meaning. An interactive fitting procedure can be stopped when no feature adds significantly more to the model, as sketched in [7].

III. Proposed Approach

A workshop was held jointly with the multimedia researchers from LIMSI credited as co-authors of this paper, that also led the MediaEval person discovery task [1]. This helped highlight some limitations in their post-hoc analysis summarized in Section II-A, and revealed a promising ground for the design of an interactive tool. Basically, decision tree learning algorithms proceed greedily, by selecting the most prevalent property at each learning stage. The need to investigate the influence of properties with a lesser importance led the researchers to a two-step approach. They first identified that two OCR (*Optical Character Recognition*)-related properties, heavily used by unsupervised multimodal systems, are prevalent in the post-hoc analysis. To focus on weaker signals, they removed these two properties from the analysis, along with the couples associated with low-entropy nodes of the tree. Remaining properties were aggregated using a DBSCAN step, and used to fit decision tree models on the selected subset of data.

Conclusions reached on the full data set, without data set tweaking, would be stronger. Also, feature discretization as performed by DBSCAN does not account for predictive power of features. Moreover, the oracle setting obfuscates individual run specificities, thus limiting the possible insight.

In this section we describe the proposed visual tool, designed to enable the post-hoc analysis task while limiting the identified caveats. As an alternative to the greedy feature selection, we use a stepwise modelling approach [5], [7]. The process starts from trivial models (i.e. only intercept terms), and features can be added and removed interactively, with classification models updated in the background. Each target is modelled independently instead of using an oracle as in [1].

A. Stepwise Multi-Target Feature Selection

The performance associated to individual runs of retrieval states can be characterized by the *rate* of the positive state in the run. In addition, the most common criterion to estimate the ability of a classification algorithm to predict both positive and negative retrieval states is the *accuracy*. However this measure is sensitive to the imbalance of the target: e.g. the accuracy equals the rate of the majority class for a model that always predicts this class. This is known as the accuracy paradox [8]. This is problematic in this paper, as we intend to compare the influence of features on predicting targets with variable rates. Hence we evaluate classification model quality with Cohen's Kappa [9], that effectively measures the probability of random agreement. The paradoxical case mentioned above then yields Kappa $= 0$ irrespectively of the class imbalance.

In this paper we use decision trees [2] (DT) and generalized linear models [10] (GLM) as classifiers. They were chosen for their low (linear) computational complexity for learning with a known set of properties. Formally, a model M learns a prediction function from a set of features S. In a stepwise context, S is initially $\emptyset$ and is augmented interactively by the user. Each feature addition triggers the update of the model so that the feature is accounted for. Let us assume we want to estimate a score that measures the improvement brought by the inclusion of feature f in the existing model M^S. The new model is then denoted as $M^{S \cup f}$. The likelihood ratio is only suited for models with similar architecture. As an alternative for models with variable architecture, McNemar's test statistic is based on the relative performance of M^S and $M^{S \cup f}$ [11]. Its null hypothesis is that predictions of these two models are not significantly different. The negative log of the p-value of this test measures the rejection strength of this hypothesis: it is therefore a meaningful score for the gain incurred by including f in the model.

The null hypothesis in [11] does not reflect if $M^{S \cup f}$ is better than M^S, i.e. only the significance of their difference is tested, and the obtained score is strictly positive. It is made relative by negating it if the feature addition or removal yields worse performance.

B. Visualization and Interactivity

Post-hoc analysis is focused on a set of targets, and statistical models that try to predict them. But in an interactive context a user has to inform his decisions about features to include in models: [3] showed that features have to be

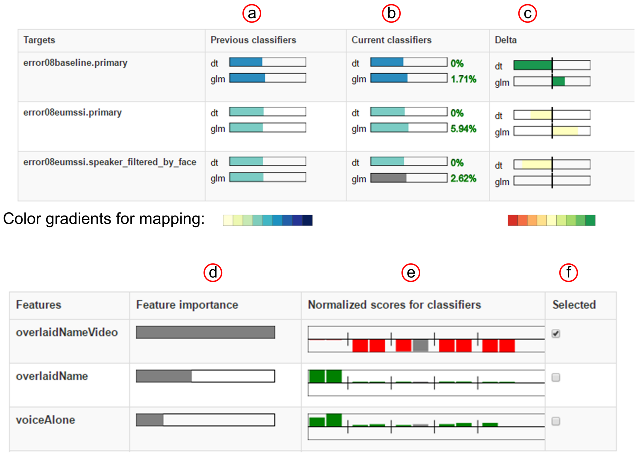

Figure 1. The target view associates *a)* rates and Kappa values to filling bar glyphs. *b)* Current models are shown along with their state preceding the latest feature selection in two distinct columns. *c)* Selecting a bar triggers the *Delta* column, that compares a reference model to all models in the same column. The reference is highlighted in gray. The feature view ranks features *d)* according to their decreasing importance. *e)* A polarized bar chart shows the positive or negative impact of modifying the feature selection state. *f)* Checkboxes let the user proceed with the interactive fitting process.

considered as first class objects in such situation. In the proposed tool, we therefore have two views: the target view and the feature view (see Figure 1).

Each target loaded in the tool is associated to a DT and a GLM. For users it is important to see both the classification rate inherent to the target, and the Kappa values of models that predict this target. As these values are independent, we opted for a filling bar design to represent each *(target, classifier)* couple (see Figure 1a), where the rate is mapped to the bar shade, and the Kappa value to the filling size. The filling design is motivated by Kappa values being bounded by 1. The *YlGnBu* quantitative color scale, taken from [12], is commonly chosen to represent positive numerical values.

To facilitate comparisons in the course of the interactive fitting process, we display two columns of models: the current models (see Figure 1b), and those that were estimated before the latest change to the set of active features (see Figure 1a). As comparing two horizontal filling bars laid out horizontally is difficult, we display the variation next to the current models. To enable comparisons within current (or previous) classifiers, clicking one of the bar glyphs reveals the *Delta* column (see Figure 1c). Normalized rate variations are now mapped to the *RdYlGr* color scale. This mapping of relative values was more natural to users.

At design stage, we had the choice between tying all classifiers to the same feature selection, or enabling different feature sets for each classifier. We selected the former option, as it favors fair model comparison, and minimizes user interactions. However this is the most computationally intensive option, i.e. approximately 5s are required for each modification of the active feature set with a naive implementation (e.g. no caching) on 4-core commodity hardware.

Features can be added or removed from the feature view. Instead of using compact glyphs such as introduced in [3], we use a ranked list design, as [4] suggested this design is appropriate for up to 100 elements in the list. The user can then simply proceed with his interactive fitting process by selecting or unselecting features, viewing the impact of his latest action, and use this information to loop back in the process. The *Importance* column is built from the maximal absolute score of the respective feature in all *(target, classifier)* couples (see Figure 1d). The *Normalized scores* column provides detailed information about the expected impact of the feature on each *(target, classifier)* couple. This enables finer analysis, e.g. if the user wants to compare the models of two specific targets. A polarized histogram shows the normalized scores reflecting the influence that would result from feature addition or removal (see Figure 1e).

IV. Case Study: Participant Runs Comparison

We demonstrate the effectiveness of our tool on real data obtained from the MediaEval multimodal person identification task, described in the introduction. Within participant submissions established on approximately 27k shots, managers of the MediaEval task have extracted 2204 *(shot ID, speaker name)* couples, so as to ensure a better balance between positive and negative retrieval states. Shots have been annotated with 23 intrinsic properties (e.g. *overlaidName*, *headFront*). Expanding categorical features to binary variables yields a total number of 42 features.

The 3 MediaEval task managers are experienced multimedia researchers, and acted as the users in this section. After a training session with the tool, the users focused on the five participant runs: the baseline system (*baseline_primary*), the primary runs of the two best performing teams (*eumssi_primary* and *ssig_primary*) and one of their contrastive runs (*eumssi_speaker_filter*, with improved face detection during speech segments, and *ssig_faunion*, with an additional audio and face clustering step).

Before starting the interactive modelling process, they browsed in the feature view, and realized that 16 features have absolutely no expected impact. Inspection showed these features are very sparse (almost always 0 or false). Then they saw a distinction between the baseline and other candidate runs clearly materialized in the feature view: the most important features (approximately 13) are either effective at explaining the error pattern of the baseline or of the candidate runs.

From the visualization, they delimited two sets. One is better at explaining baseline: *overlaidName*, *lipActivityFront*, *voiceAlone* and *headFront*. The other is better at explaining candidates: *overlaidNameVideo*, *sing*, *music*, *NbOtherSF*, *voiceOthers*, *voiceOverlapping*, *nbOtherFrontHead*, *headMove* and *headProfile*.

This confirms that beyond their balance (reflected by the rates mapped to the bar fillings), baseline and candidate runs expose two drastically different distributions. From the view of multimedia researchers, the first set of features reflects the most natural properties that would favor effective multimodal recognition: the fact that the speaker name to recognize is overlaid during the shot, or that his head is facing the camera and his lips are moving.

The second set contains features with more subtle contribution to person discovery: e.g. whether the speaker name is overlaid anywhere in the video, or multiple people are heard or seen simultaneously. A large part of the shots in the complete corpus are fairly easy to identify, and these features reflect multimodal patterns that technology developers might have wanted to take advantage of (e.g. propagation of overlaid names, multiple face tracking).

Independently of the detailed normalized scores, the users first tried to include the feature with the highest importance, *overlaidName*. This high rank was expected, as the salience of OCR-related features has already been reported in [1]. This is due to the fact that the MediaEval person discovery task is fully unsupervised, i.e. speaker name inferences can only be established by propagating names extracted by OCR and speech recognition in the training corpus.

All models then immediately become almost saturated, irrespectively of the target and model architecture. Kappas for candidate runs range from 50.85% to 52.43%, when it reaches 59.45% for *baseline_primary*. This somehow contradicts observations reported in [1], that reported that *overlaidNameVideo* yielded some improvement then. This can be an effect of unfolding the oracle run they used, or of using Kappa instead of classification accuracy.

Alternatively, the users wanted to take advantage of the initially observed feature categorization, and, starting from the trivial model, added features from the second set in a greedy fashion until saturation of the models. Features are chosen so that they influence positively as many candidate run models as possible. Completely ignoring features from the first set yields a Kappa range of 49.34% to 52.62% for the candidate runs. Hence we get comparable models with a completely disjoint set of features, that characterize preferably candidate runs (the best for *baseline* is DT then, with Kappa = 35.14%).

Saturation is much slower in the latter case, e.g. adding *overlaidNameVideo* alone yields Kappa in $[35.26, 37.73]$. Next most important feature is *sing*, that brings $[35.26, 44.16]$. We emphasize the interest of having multiple models: these two extrema are obtained when modelling EUMSSI runs either with DT of GLM architectures. Histograms show *music* has a complementary effect then.

V. Conclusion

This paper described a novel visual predictive analytics tool to support the post-hoc analysis of runs issued by participants to a multimedia person discovery task. Its motivation has been established from the review of relevant work, and direct involvement of three multimedia researchers, that provided real data resulting from the recent MediaEval person discovery task [1]. They acted as users in a case study, which provides empirical evidence of the usefulness of the tool. In particular, the visualizations facilitated the identification of two feature categories, which was not possible using the prior approach.

References

[1] J. Poignant, H. Bredin, and C. Barras, “Multimodal Person Discovery in Broadcast TV: lessons learned from MediaEval 2015,” *Multimedia Tools and Applications (submitted)*, 2016.

[2] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] J. Krause, A. Perer, and E. Bertini, “Infuse: interactive feature selection for predictive modeling of high dimensional data,” *IEEE TVCG*, vol. 20, no. 12, pp. 1614–1623, 2014.

[4] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “Lineup: Visual analysis of multi-attribute rankings,” *IEEE TVCG*, vol. 19, no. 12, pp. 2277–2286, 2013.

[5] C. A. Steed, J. Swan, T. Jankun-Kelly, and P. J. Fitzpatrick, “Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates,” in *IEEE VAST*, 2009, pp. 19–26.

[6] G. J. McLachlan, “On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture,” *J. R. Stat. Soc. Series C (Appl. Stat.)*, vol. 36, no. 3, pp. 318–324, 1987.

[7] P. Bruneau, M. Stefas, H. Bredin, J. Poignant, T. Tamisier, and C. Barras, “A visual analytics approach to finding factors improving automatic speaker identifications,” in *ACM ICMI*, 2015, pp. 323–326.

[8] F. Valverde-Albacete and C. Peláez-Moreno, “100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox,” *PloS one*, vol. 9, no. 1, 2014.

[9] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software*, vol. 33, no. 1, pp. 1–22, 2010.

[11] N. Breslow, N. Day, and J. Schlesselman, “Statistical Methods in Cancer Research,” *J. Occup. Env. Med.*, vol. 24, no. 4, pp. 255–257, 1982.

[12] M. Harrower and C. Brewer, “ColorBrewer.org: an online tool for selecting colour schemes for maps,” *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.