

Exemplar-Based Complex Features Prediction Framework

Mohamed ABOU-ZLEIKHA, Julie CARSON-BERNDSEN
CNGL, School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
mohamed.abou-zleikha@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract—Exemplars are typically defined by set of features that may have simple or complex structures. Comparing two exemplars requires a distance calculation between their features, a task which becomes more difficult when some of these features are missing. A possible solution is to predict the missing features making use of those that are known. Prediction of features is considered a hard task in machine learning and becomes more difficult when features have a complex structure and the relationship between the features is not clearly defined. This paper presents a framework for predicting complex features based on exemplar theory. The framework presented consists of two stages. The first stage is the similarity correlation stage, in which the correlation between the distance matrices of the features is calculated to determine the relationship between missing and existing features. The second stage calculates the conditional membership probability between these features using the distance matrices; this value determines the probability that for a new example not found in the dataset for which only some features are known, an exemplar with similar features to those of missing features that can be adapted to serve as appropriate features for the new example. This paper also presents a case study for the use of the framework in the context of speech synthesis. The framework is used to investigate the relationship between duration information and the syntactic and dependency trees.

Keywords-component; complex features prediction, prosody prediction, prosody text correlation, duration modelling

I. INTRODUCTION

Comparison and prediction of feature have been the focus of attention of many studies in the field of machine learning. Different techniques and methods have been applied in many domains. Rule-based, statistical based and exemplar-based techniques have been implemented and tested for different types of features [1]. An exemplar-based feature prediction method has been investigated by [2,3,4] among others. Several prediction models have been proposed and they mostly rely on categorised learning data, obtained from tasks in which subjects learn new categories of an artificial stimulus. Less work has been carried out on feature prediction when the categorisation is uncertain [5,6,7]. The methodologies employed in these approaches use the whole set of exemplars from different categories in order to predict the missing features. The results showed correlation sensitivity in the form

of statistical correlation between features that are not correlated in nature [5].

Given an example with m features, with some of which have a complex structure, the purpose is to predict a feature which is missing in this example. In an exemplar-based approach, this problem is usually solved by find the k correlated features to the missing and comparing these features of this example with the corresponding features from a set of exemplars, where the missing feature is known, and select the closest one as a predictor for the missing feature; alternatively by calculating the probability of each possible value for this feature given the values of the existing features, the missing feature can be predicted. In order to apply these approaches, the correlation between the features needs first to be calculated which can be difficult because of:

- the complexity of the structures of the features (e.g. tree, matrix), which makes defining the statistical moments of the data not always possible
- the mismatch between the dimensions of the features

Finding a solution to the correlation problem only resolves the comparison part of the task; predicting the most likely values of features has its own difficulties because the similarity function is not always an identical function, but rather, a continuous distance function, and choosing the closest exemplar to the new example does not always provide a good solution, since the mapping between these features is not always a one-to-one mapping. On the other hand, calculating the probability is not always possible because the set of possible values of the features is infinite. These issues make it difficult to apply the traditional correlation functions between the complex features, and the traditional exemplar-based prediction strategies might not be appropriate since the correlation function is not always a linear function.

The main contribution of this paper is the presentation of a framework to define an integrated strategy that can identify the correlated complex features and predict the missing ones. The conditional membership probability between these features has been proposed as a better alternative to traditional similarity functions. This probability will answer the question as to whether for a new example not found in the dataset for which only partial features are available, an exemplar with similar features can be found with the associated missing features. This

new exemplar can then be adapted to serve as a candidate for the prediction of features.

The framework uses the distance matrices instead of absolute feature values, and attempts to detect the correlated features according to these matrices, and then it calculates the probabilities that define the mapping between the correlated features. This paper also presents the use of this framework to define the relationship between the syntactic, dependency trees and duration information according to different features.

II. PREVIOUS WORK

There appear to have been only a few studies on feature prediction where categorisation is uncertain. The feature conjunction approach [5,6] has been suggested which focuses only on exemplars that have a certain feature apart from their categories, and the prediction is based on other features of these exemplars, based on the conditional probability of feature f having the value $x_i \forall i \in X$, where X are the possible values of x , given feature g having the value y . Nominal features are usually used in these experiments.

Nevertheless, the conjunction approach cannot be applied directly when features have complex structures and when the similarity function is not an identical function due to features not being nominal. An alternative approach is needed for prediction which considers the distances between the features rather than the absolute feature values.

III. EXEMPLAR-BASED FEATURES PREDICTION FRAMEWORK

The following sub-sections describe the stages of the framework for feature prediction.

The framework consists of two stages:

- 1) *calculate the distance correlation value to find the correlated features.*
- 2) *calculate the conditional membership probability to find the closest feature.*

Defining the conditional membership probability helps to answer the question as to whether, for a new example not found in the dataset for which only partial information is available, an exemplar with similar information can be found that can be adapted to serve as an appropriate predicted feature for the example. In the next two sub-sections the distance correlation and the distance-based feature prediction models are explained.

A. Feature Correlation

Finding the correlated features is an essential stage in the framework, since this relationship defines which features can be used in the prediction process. As mentioned above, finding the correlation between the features becomes difficult when the structures of the features (e.g. tree, matrix) are complex or there exists a mismatch between the dimensions of the features, making it difficult to find the dependency between these features. To solve this problem, a distance correlation [8] is used. The distance correlation is a statistical measure between two random variables, or two random vectors of values, not

necessarily having equal dimensions. This correlation is calculated using distance matrices, where the distance matrix is a matrix of $n*n$ elements containing the pairwise distances of n exemplars. The distance correlation is considered an extension of Pearson correlation [8]. Given have a feature f which is associated with a distance function dis_f , a distance matrix is calculated as follows:

$$X_{i,j} = dis_f(f_i, f_j) \quad (1)$$

where f_i is the value of feature f in the exemplar i , and the same for j . These matrices are used for correlation identification and in the calculation of the conditional membership probability. Each feature has a special distance function according to its structure and nature, but the distance functions outputs have the same type, which is a floating point.

Given the distance matrices X, Y for features f_x, f_y respectively, the centred distances are calculated for each $X_{k,l}$, where $k \in n$ and $l \in n$ are as follows:

$$A_{k,l} = X_{k,l} - X_{k.} - X_{.l} + X_{..} \quad (2)$$

where $X_{k.}$ is the k -th row mean, $X_{.l}$ is the l -th column mean, and $X_{..}$ is the grand mean of the distance matrix of the $X_{k,l}$ sample. The same calculation is done for the matrix Y and the results are saved in matrix B . The sample distance covariance is then defined as

$$dCov_n^2(X, Y) = \frac{\sum_{k,l} A_{k,l} B_{k,l}}{n^2} \quad (3)$$

and the sample distance correlation is defined as:

$$dCor_n(X, Y) = \frac{dCov_n(X, Y)}{\sqrt{dVar_n(X)dVar_n(Y)}} \quad (4)$$

where the sample distance variance is the square root of

$$dVar_n^2(X) = dCov_n(X, X) \quad (5)$$

The value of this correlation is $0 \leq dCor \leq 1$, where 0 means that the variables are independent and 1 means that there is a very strong linear relationship between the two variables. A high value for the correlation coefficient indicates that the correlation between the two features is high, which means that their similarity values are correlated. The correlation values provide an indication as to which features should be taken into consideration in next stage.

B. Conditional Membership Probability

The purpose of this stage is calculate the probability of choosing an exemplar, from the set which represents the closest $\alpha\%$ to an element i containing the f_y feature, which is also an element of the set which represents the closest $\beta\%$ to that element i containing the f_x feature. This requires the calculation of the conditional membership probability.

Given:

- Posterior feature information $X = [x_1 \dots x_n]$; where $x_i = [x_{i1} \dots x_{in}]$ is the distance vector between an example i and the rest of the data according to the defined

distance function on X . x_i is a sorted distance vector to the example i according to this feature.

- Prior feature information $Y=[y1...yn]$ where $y_i=[y_{i1}...y_{in}]$ is the distance vector between example i and the rest of the data according to the defined distance function on Y . y_i is a sorted distance vector to the example i according to this feature.

If the closest $\alpha\%$ from x_i is chosen and named ($x\alpha_i$) and the closest $\beta\%$ from y_i is chosen and named ($y\beta_i$), the question then becomes, what is the probability

$$P(z \in x\alpha_i | z \in y\beta_i) = \frac{P(z \in x\alpha_i \cap z \in y\beta_i)}{P(z \in y\beta_i)} \quad (6)$$

Figure 1 illustrates this probability.

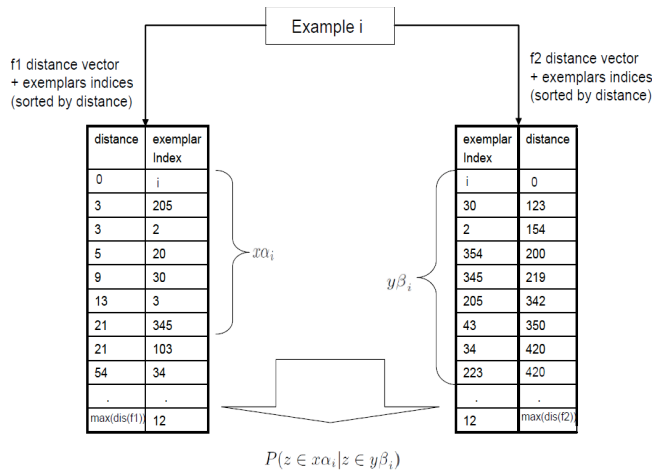


Figure 1. The conditional membership probability.

To generalise this formula for n features to be predicted from m features, the conditional membership probability becomes:

$$P(z \in \wedge x\alpha_i^v | z \in \wedge y\beta_i^w) = \frac{P(z \in \wedge x\alpha_i^v \cap z \in \wedge y\beta_i^w)}{P(z \in \wedge y\beta_i^w)} \quad (7)$$

where $\wedge x\alpha$ is calculated by taking the intersection between the indexes of the distance matrices for the v predicted features, and $\wedge y\beta$ is calculated by taking the intersection between the indexes of the distance matrices for the w existing features.

An increase in the posterior membership data leads to an increase in the conditional membership probability, but this also will lead to a decrease in the accuracy of selection. On the other hand, increasing the prior membership data leads to a reduction in the size of the intersection between the prior and posterior data and a decrease in the accuracy of the selection as well. A trade-off is needed between the membership probability and the percentage of the selected data which represents the accuracy of the similarity. The next section presents the application of the framework in the context of speech synthesis to investigate the relationship between duration features and syntactic and dependency trees. The use case that is presented is investigation the relationship between

duration features and the text features (and build the duration model according the results)

IV. CORRELATING TEXT WITH DURATION

Duration modelling plays an important role in the intelligibility and naturalness of the synthesised speech. Several studies on speech synthesis have been undertaken in order to predict the duration from text. Some of these studies investigate the relationship between text and duration information using rule-based systems, others used statistical and exemplar-based methodologies [9,10,11].

In exemplar-based duration modelling, it is necessary to find a criterion to choose an exemplar from the exemplars cloud. This criterion depends on the text features of the input. When a new input is presented, it is compared against the exemplars, and the closest exemplar (or set of exemplars) found to the new input is selected. In such models, the criteria of choosing an exemplar from the exemplars could play an important role in the model's performance. This criterion depends on the text features of the input.

In this section the previously explained framework is applied to investigate the relationship between the duration features and text features. In the next sub-section, the data used in the experiments is described.

A. Description of the Data

The CMU Arctic speech corpus is used for the study undertaken in this paper. The corpus contains 1132 utterances spoken by a US English male speaker. For each utterance, two types of information are extracted:

1) Information about the text, which includes:

- Syntactic tree (ST): which represents the syntactic structure of an utterance according to the language grammar.
- Dependency tree (DEP): which represents the grammatical dependencies between words in the utterance.

2) Information about duration, which contains:

- a) z-score of word duration
- b) z-score of syllable duration
- c) z-score of phone duration.

The syntactic and dependency trees are extracted using the Stanford parser [12,13]. The z-score which represents the deviation of the value above or below the mean, is calculated as:

$$z - score(x) = \frac{x - \mu}{\sigma} \quad (8)$$

where μ is the mean value of the duration of the unit x and σ is the standard deviation. For the text information, tree edit distance function (TED) [14] which represents the number of operations (insertions, deletions and substitutions) required to transform the tree representation of the first text to the tree representation of another is used. For the duration information, dynamic time warping (DTW) with Euclidean distance is used

to calculate the distance matrix for (a), (b) and (c) above. Due to the complexity of the features, neither statistical moments for the tree structure of the text information nor for the structure of the duration (which also have different dimensions) can be calculated. However, the similarity between the two trees is not an identical function, since it may not be possible to find an exact match but rather, a similar one. This similarity is a continuous function. The previous points could be applied on the duration features as well. The probability of a direct mapping between these two features (taking the closest exemplar according to the text information and considering its duration information as a predicted feature) is very low (about 0.02).

For the remainder of this paper, the term ST-TED will be used for the distance matrix of the syntactic tree calculated according to the tree edit distance function, and the term DEP-TED for the dependency tree calculated according to the tree edit distance. The terms ZWL, ZSL and ZPL will be used to represent the distance matrices that are calculated for the duration information on the word level (a), syllable level (b), and phones level (c) respectively.

B. Similarity Correlation

To determine the correlated features, the first stage of the framework should be applied by calculating the distance correlation between each pair of the distance matrices as describe in formula 4. Table I presents the results of applying the correlation. The results in Table I show a correlation between the similarity of the syntactic tree and dependency tree using tree edit distance function with ZWL, ZSL and ZPL.

TABLE I. DISTANCE CORRELATION BETWEEN THE SYNTACTIC TREE AND DEPENDENCY TREE USING TREE EDIT DISTANCE FUNCTION WITH ZWL, ZSL AND ZPL

	TED-ST	TED- DEP
ZWL	0.4871	0.4720
ZSL	0.4816	0.4738
ZPL	0.5419	0.5279

C. Conditional Membership Probability

The conditional membership probability as described in formula 6 is applied on the matrices which showed a high correlation value for two text information distance matrices: ST-TED, and DEP-TED, and three duration features matrices; ZWL, ZSL and ZPL. Different values for α and β were used to calculate the probability for each pair of matrices. Table II illustrates the conditional membership probability between ST-TED and ZPL. Table II illustrates that choosing 10% for the text information and 30% from the duration information gives a good accuracy. The same experiment has been carried out for the other distance matrices and Table III illustrates the probability of picking an exemplar from the 10% closest exemplars to the input for text information, and finding it in the 30% of closest exemplars in duration information.

TABLE II. THE CONDITIONAL MEMBERSHIP PROBABILITY BETWEEN ST-TED AND ZPL

		ST-TED closest data (prior)				
		10%	20%	30%	40%	50%
ZPL closest data (posterior)	10%	0.2677	0.2463	0.2275	0.211	0.1973
	20%	0.4359	0.4097	0.3838	0.3589	0.3360
	30%	0.5736	0.5470	0.5202	0.4928	0.4659
	40%	0.6834	0.6613	0.6377	0.6114	0.5847
	50%	0.7769	0.759	0.7406	0.7179	0.6937

TABLE III. THE CONDITIONAL MEMBERSHIP PROBABILITY OF PICKING AN EXEMPLAR FROM 10% CLOSEST EXEMPLARS TO THE INPUT FOR TEXT INFORMATION, AND FINDING IT IN THE 30% OF CLOSEST EXEMPLARS IN DURATION INFORMATION.

	ST-TED	DEP-TED
ZWL	0.5506	0.509
ZSL	0.5342	0.4935
ZPL	0.5736	0.5510

D. Duration Model

The level of representation (phones, syllables, words) may play a role when building a duration model. The framework supports this through the use of conditional membership probability to determine which level or combination of levels best allows duration to be predicted from the text information of exemplars. This is performed by calculating the conditional membership probability for the matrices combination with respect to the hierarchical structure (word, syllable duration) using 30% closest exemplars to an external example on the duration level and the 10% of the closest exemplars on the text level. The matrices combined are: 1) ST-TED + DEP-TED 2) ZWL+ZSL 3) ZSL + ZPL 4) ZWL + ZSL + ZPL. These combinations have been calculated by summing the normalised distances for each matrix. Table IV shows the conditional membership probabilities for the combinations. The best value is obtained for the combination of the three duration features with the combination of two distance values. The combination of phone level and syllable level features with ST-TED also gives a very close result to the best one, and with less processing.

TABLE IV. THE CONDITIONAL MEMBERSHIP PROBABILITIES FOR THE COMBINATIONS OF TEXT INFORMATION AND DURATION INFORMATION

	ST-TED	DEP-TED	ST-TED + DEP-TED
ZWL+ZSL	0.574	0.5245	0.5826
ZSL + ZPL	0.608	0.5485	0.599
ZWL+ZSL+ ZPL	0.5754	0.5596	0.609

V. DISCUSSION AND CONCLUSION

This paper has presented a framework for exemplar-based feature prediction. The framework consists of two stages: the first stage investigates the relationship between the features and the second stage defines a probability measure for this relationship which provides the basis for identifying the criteria for exemplar-based prediction. This paper focus on answering the question of whether, for a new example not in the dataset for which only part of information is available, an exemplar with similar information can be found with associated missing information which can be adapted to serve as an appropriate new model for the new example.

The framework has been used to investigate the relationship between duration information and syntactic and dependency trees. In this use case, the first part shows a correlation between z-score of words duration, z-score of syllables duration and z-score of phones duration from the side of the duration information, and the syntactic and dependency tree from the side of text information.

The results obtained from the second part of the experiment indicates a good conditional membership probabilities for the correlated data, about 0.55 for ZWL, 0.53 for ZSL and 0.57 for ZPL when 10% from the text information and 30% from the duration information are chosen. In the context of investigating the effect of the number of levels (phones, syllables, words) on the conditional membership probability in order to build a duration model, the experiment shows that better duration modelling can be obtained by using more than one level of hierarchy, where using ZSL and ZPL with ST-TED gives the best conditional membership probabilities.

At present it is difficult to compare these results directly against other approaches due to the fact that this paper uses different features and methodologies to those employed elsewhere; a comparison with other approaches will be possible when the results are incorporated into the speech synthesis system which constitutes the next step of this work; where the for a new input text, the syntactic tree is extracted, the tree edit distance between this tree and the corpus utterances syntactic trees is calculated; and then a duration template is selected and adapted (or generated) from the closest $\alpha\%$ (10% according to the result of this study) utterances to the input; as a result, an utterance structure with duration tags are generated.

The main contribution of the framework presented in this paper is that it overcomes the shortcomings of traditional exemplar selection approaches, where the probability that the best exemplar is chosen is very low and calculating the conditional probability of the feature values is not always possible.

Future work includes integrating the exemplar-based duration model into unit selection speech synthesis, and using

the correlation between the syntactic information and duration information as a criterion for selecting an exemplar from the corpus that can be used as a template for the input of the synthesiser after adapting it according to the input.

ACKNOWLEDGMENT

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngli.ie) at University College Dublin, Ireland. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors would like to thank Noor Shaker for fruitful discussions on machine learning and Dr. Peter Cahill and Dr. Fred Cummins for valuable discussions on duration modelling.

REFERENCES

- [1] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics), 1st ed. Springer, Oct. 2007.
- [2] R. Nosofsky, "Attention, similarity, and the identification-categorization relationship." *Journal of Experimental Psychology: General*, vol. 115, 1986.
- [3] G. Storms, P. De Boeck, and W. Ruts, "Prototype and exemplar-based information in natural language categories." *Journal of Memory and Language*, vol. 42, 2000.
- [4] K. Holyoak, H. Lee, and H. Lu, "Analogical and category-based inference: A theoretical integration with bayesian causal models." *Journal of Experimental Psychology: General*, vol. 139, no. 4, p. 702, 2010.
- [5] B. Hayes, C. Ruthven, and B. Newell, "Inferring properties when categorization is uncertain: A feature-conjunction account," in *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 2007.
- [6] C. Papadopoulos, B. Hayes, and B. Newell, "Non-categorical approaches to property induction with uncertain categories," in *Proceedings of the 31st annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2009.
- [7] G. Murphy and B. Ross, "Uncertainty in category-based induction: When do people integrate across categories?." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 36, 2010.
- [8] G. Szekely and M. Rizzo, "Brownian distance covariance," *The annals of applied statistics*, vol. 3, 2009.
- [9] D.H. Klatt, "Interaction between two factors that influence vowel duration" *The Journal of the Acoustical Society of America*, 1973.
- [10] W.N. Campbell, "Syllable-based segmental duration" *Talking machines: Theories, models, and designs*, 1992.
- [11] O. Goubanova, S. King, "Predicting consonant duration with Bayesian belief networks," *Proc. of the Interspeech*, 2005.
- [12] D. Klein and C. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003.
- [13] M. De Marneffe, B. MacCartney, and C. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, 2006.
- [14] E. Demaine, S. Mozes, B. Rossman, and O. Weimann, "An optimal decomposition algorithm for tree edit distance," *Automata, languages and programming*, 2007.