Similarity Searches in Heterogeneous Feature Spaces

TILMANN STEINBERG, FILLIA S. MAKEDON, JAMES C. FORD, YUHANG WANG
Department of Computer Science
Dartmouth College
Hanover, NH 03755
UNITED STATES OF AMERICA

Abstract: Correlating event streams or development paths of observed behavior that involves disparate types of data is a common problem in many applications including biomedical and clinical diagnosis systems. We present a new problem formulation of specifying feature space with heterogeneous dissimilarity measures, and trying to find similar time series given these (expert) user-specified heterogeneities, both within the same feature and as combinations across multiple features. By allowing domain experts to describe their feature spaces more accurately in this fashion, query matches are better suited to the domain experts' needs. The presented work augments the existing research of finding local similarity areas and overall patterns in time series data.

Key-Words: database queries, dissimilarity measures, prior knowledge

1 Introduction

We consider the problem of similarity searches over sequences of observations, rather than the more usual problem of measuring the similarity between individual observations. This problem arises naturally in medicine, where there is considerable interest in tracking and monitoring all possible physical signs that might be indicative of disease or other disorders, especially when assessing treatments or monitoring degenerative disorders. The subject scan data are collected at somewhat regular intervals, yielding a rich and diverse set of event streams encompassing very heterogeneous data types: for example, colleting subject data for tracking brain diseases may involve data from PET, MRI, and spectroscopy in addition to clinical evaluation data. Rather than comparing individial states of the disease, the clinician is more interested in patterns and trends, allowing for some variation in certain data points based on previous knowledge; in other words, we want to specify some already existing correlations or similarities within a set of subspaces of the feature space that the collected event streams inhabit, and then apply existing techniques for finding new correlations or similarities given the model specified by the domain expert.

This approach augments previous work in similarity search in that it creates a more specific model of the query problem, in that (a) additional information can be given by the domain expert that the existing techniques can and should incorporate, and (b) the feature space is represented with area-specific dissimilarity measures that are independent
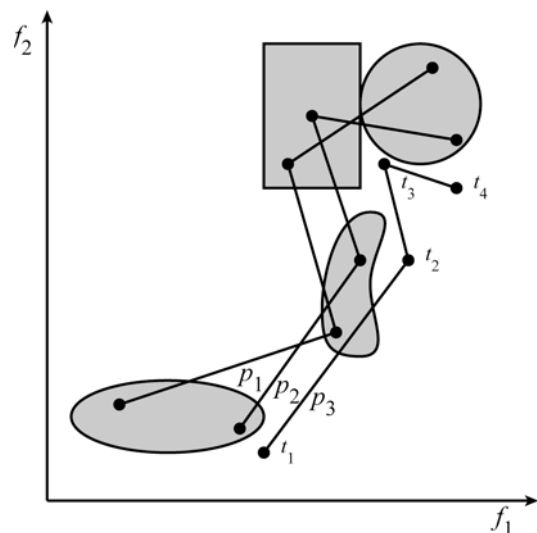
Fig. 1: Examples of paths in a feature space with two features and previously defined similarity areas (shaded). Points inside each shaded area are considered to be similar regardless of distance between them. As a result, paths p_1 and p_2 are similar to each other, but path p_3 is dissimilar from the first two.

of the actually collected data. Furthermore, this approach expands on the idea of constraint-based specification of prior knowledge by providing a more general, area-based formulation of known or assumed correlations between data points.

As an example, a clinician may want to ignore the exact location of objects found in a particular brain region while still distinguishing objects outside that region by their location, but only if a certain condition holds. One option would be to remap the objects that fulfull both the spatial and

conditional requirements, but that may not be desirable or possible. The alternative proposed in this work is to define a local dissimilarity measure for the subspace of the feature space that reflects the desired exception.

1.1 Related Work

Previous work on high dimensional similarity searches has taken into account that practical applications of high-dimensional feature spaces can have local heterogeneities. The primary solution to this problem has been to split up the feature space into locally manageable areas, or to find subspaces where local rules can apply.

Chakrabarti and Mehrotra [1] developed a technique for applying existing local dimensionality reduction methods that takes into account local correlations in the data, noting that global dimensionality reductions can have problems using local correlations, by either yielding wrong results or not using the local correlations to their full advantage (for example, data is correlated along two independent axes); and an indexing structure using the local correlations to support range and nearest-neighbor queries.

Puuronen et al. [2] use strategic splitting of the feature space to identify the best feature subset for each instance, using decision trees with local feature selections.

Atkeson et al. [3] survey methods for optimizing queries to take only local data into account, within the context of machine learning.

Apte et al. [4] measure the degree of dissimilarity in order to split the feature space into regions with distinct characteristics.

Other works for similarity searches for time-series data in multidimensional feature spaces include:

Vlachos et al. [5] describe non-metric similarity functions based on Longest Common Subsequence matching techniques for object trajectories in two- and three-dimensional space.

Gionis et al. [6] reduce dimensionality by applying "locally-sensitive" hashing functions to points in feature space, grouping those points that are close within the feature space.

Keough et al. [7] similarly create an approximation of the original data by replacing actual time series data with simpler series of mean point and end point tuples that can then be used for indexing.

Previous works on applying prior knowledge to data sets in order to improve results include:

Gordon [8] surveys methods to specify prior knowledge in the form of constrained classification where class membership of objects is based on similarity.

Klein et al. [9] aim to take spatial clusters into account when adding constraints, noting that specification of spatial constraints outperforms instance-based constraint clustering.

2 Problem Formulation

In the abstract representation of the problem described above, we have a high-dimensional feature space in which each dimension (or set of dimensions) represents a different feature (such as physical signs or clinical data in a medical application). These features may be very heterogeneous in nature as far as their interpretation by the domain expert/data owner is concerned. Existing/Known correlations between features are represented as locally defined dissimilarity measures specified by the domain expert, and can apply to single features as well as multiple features in linear or non-linear combinations. For measuring the general dissimilarity of two points in feature space, we need to define a useful and effective means of combining the applicable local dissimilarity measures.

2.1 Local Dissimilarity Measures

Existing work has focussed on finding parts of feature spaces which behave differently than the rest, but there are applications in which such behavior is known in advance (e.g. in a medical application, it may be sufficient to know that tumors or lesions appear in particular brain regions, the exact location is irrelevant). To pass this information on to existing techniques, we propose to encode the different behavior as a local dissimilarity measure, which defines similarity for points in a subspace of the feature space. The goal is to provide users (e.g. clinicians) with a simple interface to specify these encodings either ahead of time or interactively during queries, reflecting both prior knowledge and insights gained during application on the actual data.

2.2 Global Dissimilarity Measure

Finding similar paths within our feature space is based on two steps: we compute the similarity between two individual points (point-wise dissimilarity) and between two time-series of points, or paths in feature space (path-wise dissimiliarity).

Point-wise dissimiliarity needs to take into account all applicable local dissimilarities, i.e. we need to modify existing techniques (for calculating similarity, e.g. Euclidian distance) with a lookup to check if a given pair of points is within an area for which a local dissimilarity measure is defined.

Path-wise dissimilarity is computed by finding the smallest dissimililarity between the points; the problem with existing techniques is that they assume a uniform global dissimilarity measure, which is lacking in our model: we need to compensate for the fact that some point combinations may have a local dissimilarity.

3 Problem Solution

The observed data are quantized as features f_i that form a *feature space* $F = (f_1, f_2, …, f_n)$. For each subject s, the time series of observed data translate into a *path* $P_s = (p_1, p_2, …, p_k)$ with each point p_j from F representing the observed data at a time t_j. Interpretation of values of each feature is domain-specific; for each feature, we define $d_f(x, y)$ as the *feature-specific dissimilarity measure* (for two values x and y from feature f). The *global dissimilarity measure* $d(x, y)$ gives the distance for two points x and y from the feature space F.

We define *local dissimilarity measures* $d_A(x, y \in F)$ on an area $A \subset F$, where $1 \leq \dim(A) \leq n = \dim(F)$. For each local dissimilarity measure, we store a mapping from the area A to the function d_A.

We stipulate that there is an ordering, represented as a directed acyclic graph, of these mappings (in case of conflict).

We want to (a) quickly find applicable mapping (allow for overlap: mapping A for region R, mapping B for subregion of R), hence use some sort of spatial indexing (e.g. R-tree); but also (b) store mapping only once, and be able to update it effectively (only one update). A simple solution is to keep a list of the maps and only store references to them in a spatial (R-)tree.

3.1 Representation of Local Dissimilarity Measures

The simplest type of heterogeneous dissimilarity measure is to define a distance for point pairs from a range of values (given by the interval $[a, b]$) in one feature:

$$\hat{d}(x, y) = \begin{cases} h(x, y) \text{ if } x, y \text{ in } [a, b] \\ d(x, y) \text{ otherwise} \end{cases} \quad (1)$$

where d is the standard dissimilarity measure for the feature and h is the local dissimilarity measure.

For features with discrete values (e.g. location of voxels) we can also use sets of values for which local function applies, e.g. in form of a bitmap. This is also true when extending the area to multiple features: e.g. the definition for the interior of a 3D object (e.g. a brain region of interest) can be given as bitmap B (which can be the result of a segmentation, or taken from a defined standard):

$$\hat{d}(x, y) = \begin{cases} h(x, y) \text{ if } B_{x,y} = 1 \\ d(x, y) \text{ otherwise} \end{cases} \quad (2)$$

(Here, x, y are multi-dimensional points.)

The above definitions yield mappings that are easily located in the feature space via their bounding boxes. More complicated definitions are possible by specifying conditions for point pairs as functions, however for such definitions the bounding box will need to be computed and stored separately.

3.2 Combining Local Dissimilarity Measures for Measuring Global Dissimilarity

For the combination of dissimilarity measures for different features, we can apply a simple solution by taking a linear combination of individual distances, with weights/parameters again specified by the domain expert to represent his or her interpretation of the data features. However, this has the potential of making paths dissimilar if there is a single outlier, which may not be accurate depending on the location of the outlier. One way to circumvent this is to use a combination function that de-emphasizes single large differences in the presence of numerous minor differences (e.g. through normalizing or by considering only k most similar features, where $k<n$).

3.3 Sample Application

Our work is motivated by applications in the medical field, where we want to analyze the behavior of objects obtained from image scans over semi-regular time intervals. Each object has a diverse set of features and associated data; we are interested in both the development of each object and the existence of objects with similar behavior.

In particular, the observed features include multiple locations, sizes, principal axes, basic shape descriptors (i.e. elongated vs. spherical), and

imaging properties (e.g. average intensity). One particular analysis tries to determine if there is a common development from (small) spherical shapes to (larger) elongated shapes in a particular area, and where this development correlates with associated data (such as drug treatment). In this problem, we can define local dissimilarity measures to group all objects in the area of interest, and to distinguish spherical from elongated objects. We would then query for object sequences that are located in the area of interest and progress from the spherical feature to the elongated feature.

In another (medical) application, researchers have found a strong correlation between the condition of interest (chance of development of heart tumors) and peaks in spectorscopy maps for particular variables that are relative to other variables. Using our approach, data points showing the target peaks are defined to be "close", regardless of actual peak values since we are interested only in the presence of these peaks.

4 Conclusion and Future Work

We have formulated a new approach to problem of searching for similarities in a feature space that describes multiple event streams and includes heterogeneous dissimilarity measures based on feature values and property combinations of some of the features within specified areas. The advantage of this approach is that domain experts can represent their understanding of the data independently from the collected data, in the form of subspaces of the feature space and the desired dissimilarity measures defined for each subspace. This representation is more flexible than previous approaches in that it offers more control over the exact type of similarity between different points. However, for more complex definitions, it may be more difficult to efficiently determine whether a given pair of objects is affected by a local dissimilarity.

One possible application we see for the described approach is denoising or detangling feature spaces by replacing noisy intervals (path segments) with similar data that has been either pre-set by the domain expert or chosen from a particular sample. Path segments located in previously identified, "noisy" areas are considered similar, and replaced with the pre-set or sample similar data.

For future work, we are looking at effective methods to create user interfaces for specifying standard and complex arrangements of mappings from subspaces or areas to dissimilarity measures; create optimal data structures for efficient query

service (priority graph to deal with overlaps, spatial data structure for quickly finding appropriate dissimilarity measure); and allow for more complicated encoding of expert knowledge. Finally, we are looking into adopting existing indexing methods that can be modified to work with the heterogeneous feature space.

References:
[1] Kaushik Chakrabarti, Sharad Mehrotra, "Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Space", *Proceedings of the 26th VLDB Conference,* Cairo, Egypt, 2000.

[2] Seppo Puuronen, Alexey Tsymbal, Iryna Skrypnyk, "Advanced Local Feature Selection in Medical Diagnostics", *13th IEEE Symposium on Computer-Based Medical Systems* (CBMS'00), June 23-24, 2000, Houston, TX.

[3] Christoper G. Atkeson, Andrew W. Moore, Stefan Schaal, "Locally Weighted Learning", *Artificial Intelligence Review,* Vol. 11, Ns. 1-5, 1997, pp. 11-73.

[4] C. Apte, S.J. Hong, J. Hosting, J. Lepre, E. Pednault, B. Rosen, "Decomposition of Heterogeneous Classification Problems", *Intelligent Data Analysis,* 1998.

[5] Michail Vlachos, George Kollios, Dimitrios Gunopulos, "Discovering Similar Multidimensional Trajectories", *Proceedings of 18th ICDE,* p673-684, 2002, San Jose, CA.

[6] Aristides Gionis, Piotr Indyk, Rajeev Motwani, "Similarity Search in High Dimensions via Hashing", *Proceedings of the 25th VLDB Conference,* Edinburgh, Scotland, 1999.

[7] Eamonn Keogh, Kaushik Chakrabarti, Sharad Mehrotra, Michael Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases", *Proceedings of ACM SIGMOD Conference on Management of Data,* pp. 151-162, May 2001, Santa Barbara, CA.

[8] A. D. Gordon, "A Survey of Constrained Classification", *Computational Statistics & Data Analysis,* Vol. 21, 1996, pp17-29.

[9] Dan Klein, Sepandar D. Kamvar, Christopher D. Manning, "From Instance-Level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering", *Proceedings of the Nineteeth International Conference on Machine Learning,* July 2002, Sydney, Australia.