Unsupervised Clustering of Multidimensional Distributions using Earth Mover Distance

D. Applegate  
david@research.att.com
T. Dasu  
tamr@research.att.com
S. Krishnan  
krishnas@research.att.com
S. Urbanek  
urbanek@research.att.com

ABSTRACT

Multidimensional distributions are often used in data mining to describe and summarize different features of large datasets. It is natural to look for distinct classes in such datasets by clustering the data. A common approach entails the use of methods like k-means clustering. However, the k-means method inherently relies on the Euclidean metric in the embedded space and does not account for additional topology underlying the distribution.

In this paper, we propose using Earth Mover Distance (EMD) to compare multidimensional distributions. For a n-bin histogram, the EMD is based on a solution to the transportation problem with time complexity \(O(n^3 \log n)\). To mitigate the high computational cost of EMD, we propose an approximation that reduces the cost to linear time.

Other notions of distances such as the information theoretic Kullback-Leibler divergence and statistical \(\chi^2\) distance, account only for the correspondence between bins with the same index, and do not use information across bins, and are sensitive to bin size. A cross-bin distance measure like EMD is not affected by binning differences and meaningfully matches the perceptual notion of “nearness”.

Our technique is simple, efficient and practical for clustering distributions. We demonstrate the use of EMD on a practical application of clustering over 400,000 anonymous mobility usage patterns which are defined as distributions over a manifold. EMD allows us to represent inherent relationships in this space. We show that EMD allows us to successfully cluster even sparse signatures and we compare the results with other clustering methods. Given the large size of our dataset a fast approximation is crucial for this application.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Data Mining

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD 2011 San Diego, CA
Copyright 20XX ACM X-XXXXXX-XX-X/XX/XX ...$10.00.

General Terms

Algorithms

Keywords

Earth mover distance (EMD), clustering, signatures, distributions

1. INTRODUCTION

Massive data sets are often summarized using signatures that capture important characteristics of the data. Signatures are compact representations that can be used for further mining tasks such as clustering and anomaly detection. For example, Cortes et al [3] discuss the use of signatures for mining massive telecommunications data to find communities of interest, and for detecting fraud. In [18], Wong et al demonstrate the use of data signatures for visualizing complex scientific data sets.

In many instances, signatures are multidimensional distributions that represent the distribution of probability mass across a manifold. An important data mining task involves measuring distances between such distributions, which are then used for subsequent tasks such as clustering, change detection or alerting. For instance, clustering usage signatures of telecommunications customers identifies canonical patterns of usage that can be used to provide better service and for network load balancing tasks.

While clustering algorithms such as k-means are frequently used to cluster distributions, it can lead to misleading results. Consider the following example where a population of cell phone users consists of users with different types of usage behavior. Figure 1 shows examples of different usage signatures. We defer a detailed explanation to Section 5.1, but for now let us treat them as heat maps that show the density of usage for voice calls and text messages aggregated over two months across “hour of the day” (rows in the heat map) by “day of the week” (columns). The signature in Figure 1 (left) is representative of users with a moderately high volume of calls and texts, distributed across the bins corresponding to working days. The signature in Figure 1 (right), on the other hand, is representative of users with lower usage. Such users generate very few calls or text messages, their signatures are quite sparse, typically consisting of isolated spikes. Any method that relies on a distance that treats bins separately, is bound to fail on such data.

To demonstrate, we applied a k-means clustering algorithm to a sample of 5,000 signatures from the bottom quartile of the usage distribution. The centers of the resulting 8 clus-
1.1 Distance Between Distributions

There are numerous statistical distances that measure the distance between distributions. Many of them are not true metrics because they violate one or more of the three requirements of a metric, that of non-negativity, symmetry and triangular inequality. We briefly describe some well-known metrics because they violate one or more of the three requirements.

1.1.1 Triangular Inequality

We briefly describe some well-known metrics because they violate one or more of the three requirements of the triangular inequality. We therefore do not consider any meaningful information. This is an example of a real-world application where \(k\)-means would result in arbitrary clusters.

The reason for such odd clusters is that the \(k\)-means method inherently relies on the Euclidean metric in the embedded space and does not take into account any additional topology of the bins. Distributions lie on the probability simplex, and do not use information across bins, and are sensitive to bin size. A binning that is too fine will place similar features in different bins which will never be matched. EMD is a cross-bin distance and is not affected by binning differences.

Figure 2: Cluster centers of low voice usage signatures using \(k\)-means clustering for \(k = 8\). The bold number on top represents cluster number (id), top bar represents relative cluster size. Regardless of the value of \(k\), the number of clusters, the method is unable to separate any clusters other than individual bins of the histogram. The location of the bins is random, based entirely on the initial cluster seeds.

Figure 1: Two samples of a mobile usage distributions created by binning usage by hour of day, day of the week and type of activity. Color scale of the counts in bins ranges from light yellow (zero) to red (high density). Columns of each matrix correspond to days of week, rows correspond to hours of day. Left: Moderately high usage, mostly during work hours on working days. Right: Sparse usage. For privacy reasons the samples do not match any actual, individual user’s signature in the data.

The distribution \(Q\). This class of distributions is known by the names Csiszar’s \(f\)-divergence \([4]\), Csiszar-Morimoto divergences or Ali-Silvey family of distances \([1]\). Depending on the form of \(f\), distances in this family include the Kullback-Leibler distance, Hellinger distance, total variational divergence, \(\chi^2\) divergence and alpha-divergence. Gibbs et al \([5]\) provide an overview of choosing appropriate probability metrics, including whether the metric has an interpretation with respect to the problem at hand. This is a critical, since the results of clustering are heavily dependent on the choice of the metric. Liese et al \([11]\) describe a novel approach characterizing distances and their relationships, and the derivation of new properties with respect to desirable statistical properties like sufficiency.

Other popular distributional distances include Jensen Shannon divergence which is a symmetrized version of KL-divergence, Bhattachrya distance used to measure separability of classes, and Wasserstein’s metric.

In this paper, we focus on Wasserstein’s metric, popularized in computer vision and image retrieval community by the name Earth Mover Distance or EMD. The major drawback of the Kullback-Leibler divergence and statistical \(\chi^2\) distance is that they account only for the correspondence between bins with the same index, and do not use information across bins, and are sensitive to bin size. A binning that is too coarse will not have sufficient discriminative power, while a binning that is too fine will place similar features in different bins which will never be matched. EMD is a cross-bin distance and is not affected by binning differences.
It meaningfully matches the perceptual notion of “nearness” and hence its popularity in image matching and retrieval.

EMD has not found widespread application outside the computer vision community due to its prohibitive computational cost. A lone exception has been the work of Zhou et al. [19], who propose the use of Mallows distance for comparing clusters.

By generalizing the wavelet approximation proposed in [16] to any general manifold we create a fast approximation to EMD that is applicable in any general domain. This allows us to apply it specifically to signatures that are defined on a probability manifold, not limited to flat feature spaces.

The rest of the paper is organized as follows. We define the problem, outline our approach and describe our contributions in Section 2. In Section 3 we discuss the Earth Mover Distance and its relationship to other distances like distributions in Section 2. In Section 3 we discuss the Earth Mover Distance and its relationship to other distances like Mallows distance. Section ?? describes the algorithm and a highly efficient but general approximation that makes EMD feasible on very large data. In Section 5, we conduct experiments on a massive real-world application involving anonymized mobility data. Finally, we present our conclusions and outline directions for future research in Section 6.

2. OUR APPROACH

In this paper, we address the problem of clustering distributional signatures derived from massive multidimensional data. Extant approaches typically rely on methods such as k-means that are based on the Euclidean distance. However, signatures are not mere points in a metric space, they are defined over a distinct topology of the underlying distribution. While known distributional divergences can be used for clustering distributions, the advantages of EMD discussed in Section 3 make it an ideal candidate for clustering distributional signatures. EMD has not found widespread use due to its computational cost. Our approximation makes it feasible, and to the best of our knowledge, ours is the first paper in the KDD community that uses EMD for clustering multi-dimensional distributions.

We make three contributions.

• While wavelet approximations to EMD have been proposed for Euclidian spaces, we generalize it to any manifold, thus making it widely applicable to topologies that include probability manifolds. This novel generalization allows us to apply EMD to distributional signatures, and on a previously unfeasible scale.

• use the Earth Mover Distance for the unsupervised clustering of multi-dimensional data distributions, and

• demonstrate the efficacy of the method and showcase its properties using a compelling real world application of mobility data patterns.

We start with a brief introduction to EMD in the next section.

3. EARTH MOVER DISTANCE

Rubner et al. [14, 15] introduced EMD to the computer vision community as a technique for image retrieval. It is also known as the Mallows distance [10] or Wasserstein distance in the statistical literature. EMD is a distance measure between two distributions. We start by introducing some terminology to define EMD.

Distributions can generally be described as functions over an underlying (feature) domain. In the rest of the paper, we will assume that multidimensional distributions are defined as histograms by partitioning the underlying domain into bins with a mass associated with each bin. Defining a distance between two distributions, first, requires a notion of distance between points in the underlying domain defining the distributions. This distance is called the ground distance.

EMD is a natural and intuitive metric between histograms if we think of them as piles of sand sitting on the ground (underlying domain). Each grain of sand is an observed sample. To quantify the difference between two distributions, we measure how far the grains of sand have to be moved so that the two distributions coincide exactly. EMD is the minimal total ground distance travelled weighted by the amount of sand moved (called flow). For perceptually meaningful ground distances, EMD agrees with perceptual dissimilarity better than other measures [15, 16]. Further, if the ground distance is a metric and the distributions all have the same mass (this is true in the case of probability distributions), EMD is a metric as well. There are several advantages of using EMD over other distribution dissimilarity measures. For example, it does not suffer from arbitrary quantization problems due to rigid binning strategies. Therefore, it is robust to errors in transformation that take raw data into feature space.

We provide a formal definition for EMD by using the same notation as proposed in Rubner et al. [15]. Let \( \mathcal{F} \) be the underlying (feature) space over which the distributions are defined, endowed with a distance function \( d : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}^+ \). Rubner et al. describe a distribution using signatures - a sparse representation of a histogram. Specifically, a signature is defined as a vector of tuples, \( P = \{< p_1, w_{p_1} >, < p_2, w_{p_2} >, \ldots, < p_m, w_{p_m} > \} \), where \( p \in \mathcal{F} \) is a point in the feature space and \( w_p \) is the weight. Let \( P \) and \( Q \) be two distributions of the form

\[
P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \ldots, (p_m, w_{p_m})\} \quad (2)
\]

\[
Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \ldots, (q_n, w_{q_n})\} \quad (3)
\]

and let \( D = \{d_{ij}\} \) be the ground distance between points \( p_i \) and \( q_j \). We want to find a flow, \( F = \{f_{ij}\} \), where \( f_{ij} \) is the flow between points \( p_i \) and \( q_j \), that minimizes the overall cost

\[
W(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \quad (4)
\]

The flow \( f_{ij} \) must satisfy the following constraints:

\[
f_{ij} \geq 0, \quad \forall i \in 1, \ldots, m, j \in 1, \ldots, n \quad (5)
\]

\[
\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, \quad \forall i \in 1, \ldots, m \quad (6)
\]

\[
\sum_{i=1}^{m} f_{ij} \leq w_{q_j}, \quad \forall j \in 1, \ldots, n \quad (7)
\]

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min \left( \sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j} \right) \quad (8)
\]

The first constraint limits the flow to move from \( P \) to \( Q \) and not the other way around. The next two constraints limit the flow from exceeding the weights in the corresponding bins, and the final constraint forces us to maximize the
flow as much as possible. Let \( f^* \) be the optimal flow that minimizes Equation 4. Then, the Earth Mover Distance is defined to be

\[
EMD(P, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} f^*_{ij} d_{ij} 
\]

(9)

The optimization problem above can be formulated as an uncapacitated minimum cost flow problem and can be solved optimally by using Orlin’s algorithm [9] in \( O(n^3 \log n) \), where \( n \) is the number of bins in the two histograms. The biggest drawback of using EMD for clustering distributions is its high computational complexity. When dealing with large datasets (with each distribution potentially having a large number of bins) as in our case, the exact computation of EMD becomes impractical. To alleviate this problem, we propose a modified version of an algorithm proposed by Shirdhonkar et al. [16] that works on arbitrary ground spaces.

![Figure 3: Cluster centers of the same low voice usage signature data as Figure 2, but using hierarchical clustering with EMD. Even with very sparse signatures, EMD is capable of separating them into coherent groups such as weekday (Clusters 1-5) weekend (Clusters 6-8), or time of the day (Clusters 2,3).](image)

4. APPROXIMATION TO EMD

In order to mitigate the high computational cost of EMD, several approximation schemes have been proposed in the literature. Ling and Olada [12] empirically showed that EMD could be computed in \( O(n^2) \) time if an \( L_1 \) ground distance is used instead of the usual Euclidean distance. Indyk and Thaper [7] use a randomized multiscale embedding of histograms into a space equipped with the \( L_1 \) norm. The multiscale hierarchy is obtained by a series of random shifting and dyadic merging of bins. The histogram levels are weighted by powers of 2, with more weight at the coarser levels. They show that the \( L_1 \) norm computed in this space, averaged over all random shifts, is equivalent to the EMD. Their approach runs in roughly \( O(ndk) \) time, where \( d \) is the histogram dimension and \( k \) is the number of times randomized embedding is performed. Shirdhonkar and Jacobs [16] present a linear time algorithm for approximating the EMD for low dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram. The weighted wavelet transform is computed in time linear in the number of histogram bins. One main restriction in the methods of Indyk et al. and Shirdhonkar et al. is that the underlying (feature) space must be flat (for example, Euclidean).

In the context of our work, the distributions are embedded in a manifold-like space. In these cases, the above approximation strategies are not applicable. Inspired by the work of Shirdhonkar et al., we modify their approach to extend it to the case where manifolds can be the underlying space. We model manifolds as weighted graphs and apply wavelets on these graphs using spectral graph theory [6].

We start our description of the approximate EMD computation by describing the theory behind the approach of Shirdhonkar et al. The connection becomes apparent when we look at the continuous version of the EMD problem known as the Kantorovich-Rubinstein transshipment problem. Let \( S \subset \mathbb{R}^d \) be a compact space, and let \( P_1 \) and \( P_2 \) be probability distributions with densities \( p_1 \) and \( p_2 \), respectively, defined on \( S \), endowed with a continuous cost function \( d : S \times S \rightarrow \mathbb{R}^+ \). Let \( 0 < s \leq 1 \) be a fixed parameter. Then the transshipment problem is to find

\[
\mu_d = \inf_q \int d(x, y) q(x, y) dxdy,
\]

where the infimum is over all joint probability densities \( q \) on \( S \times S \). \( q \) is analogous to the flow in the discrete setting. The joint distribution must satisfy a mass conservation constraint of the form

\[
p_1(u) - p_2(u) = \int q(u, y) dy - \int q(x, u) dx
\]

Here \( p := p_1 - p_2 \) is the difference distribution with the property that \( \int p = 0 \). The Kantorovich-Rubinstein theorem states that the above problem admits a dual representation

\[
\mu_d = \sup_f \int f(x)p(x)dx
\]

(10)

The supremum is over all bounded continuous functions \( f \) on \( S \) satisfying the order \( s \) Hölder continuity condition \( f(x) - f(y) \leq d(x, y)^s \), for all \( x, y \in S \). In the special case of \( S = \mathbb{R}^d \), and \( d(x, y) = \| x - y \| \), Shirdhonkar et al. observed that the dual objective function is the maximum inner product between the potential function and the difference histogram, which can be easily represented in the wavelet domain (since orthonormal wavelets preserve inner products). Their formulation in the wavelet domain has the explicit solution:

\[
\mu^*(p_1, p_2) = \sum_{\lambda} 2^{-j(1+n/2)}|p_\lambda|,
\]

(11)

where \( p \) is the \( n \)-dimensional difference histogram and \( p_\lambda \) are its wavelet coefficients. The main result of their paper states that, depending on the wavelet used, there exist constants \( C_L \) and \( U \) such that \( C_L \mu^* \leq \mu_d \leq U \mu^* \). We refer the reader to their paper for more details.
Diffusion wavelets [13, 17] generalize classical wavelets, allowing for multiscale analysis on general structures, such as manifolds, graphs and point clouds in Euclidean space. They allow us to perform signal processing tasks on functions on these spaces. We are interested in preserving dot products; hence we rely on orthonormal wavelets. The diffusion wavelets were designed specifically to be orthonormal applicable to smooth manifolds as well as graphs. Their construction interacts with the underlying graph or manifold space through repeated applications of a diffusion operator \( T \) (like the heat operator on graphs). The orthogonality of the wavelets is achieved by running a localized orthogonalization procedure after applying dyadic powers of \( T \) at each scale to yield nested approximation spaces. Wavelets are then produced by locally orthogonalizing vectors spanning the difference of these approximation spaces.

\[
\sum |u_i - v_i|_1
\]

Figure 4: Illustration of the EMD approximation process

We generate a weighted graph to reflect the underlying topology of our space (a product space of a torus and real line), and perform the wavelet computation once. This wavelet replaces the simpler wavelets used by Shirdhonkar et al. All the results directly extend to this wavelet except for the specific values of \( C_L \) and \( C_U \). The rest of the computation remains exactly the same. In order to perform the approximate computation, we compute the weighted wavelet transform (as in Equation 11) of each distribution independently. Once the coefficients are obtained as a sparse vector, the approximate EMD is essentially the \( L_1 \) norm of the difference of transform coefficients. Figure 4 illustrates the process of approximating EMD using wavelet transforms on a simple domain and the main steps of the algorithm are described in Algorithm 1.

Algorithm 1: Approximate EMD\((p_1, p_2)\)

**Input:** Two distributions \( p_1 \) and \( p_2 \) over a compact space \( S \subseteq \mathbb{R}^d \)

**Output:** ApproxEMD\((p_1, p_2)\)

Discretize the compact space \( S \) by sampling bin locations

Construct a weighted graph \( G \) reflecting manifold metric

Construct diffusion wavelets for \( G \) (Maggioni et al. [13])

Let \( wd_1 (wd_2) \) be the weighted diffusion wavelet descriptor (Eq. 11) of \( p_1 (p_2) \)

\[
D^* = L_1\text{-dist}(wd_1, wd_2)
\]

return \( D^* \)

4.1 Accuracy and Performance of Approximate EMD

We compare the results of exact EMD and our approximation on a simulated example and real data. For the simulation, we generate uniformly distributed vectors of size \( 10^3 \) using a sparsity factor \( \rho \). From this vector we generate two histograms: one with all positive entries and one with absolute values of only the negative entries. We then compute the EMD between those histograms using exact computation and our approximation. We repeat this process 1000 times varying \( \rho \) linearly between 0 and 1. The resulting scatterplot of the exact values vs approximations is shown in Figure 5(a).

We also used our real data from the application described in Section 5.1 to perform the same comparison. Here, we calculated the EMD from a random non-sparse signature to 5,000 random signatures. The result is shown in Figure 5(b). As we can see, there is a strong correlation between the two measures.

5. A REAL WORLD APPLICATION

In this section, we demonstrate the use of EMD based clustering of signatures of cell phone usage data and compare conventional clustering of distributions based on Euclidean distances with clusters derived using the EMD. We hypothesize that while \( k \)-means might provide a reasonable approximation in situations where the signature distributions are fairly smooth, it will struggle when the distributions have spikes e.g. when the density is concentrated in a few bins such as when the signatures are sparse as shown in Figure 1 (right). We have already discussed in Section 1 that \( k \)-means results in arbitrary clusters determined entirely by the initial seeds, illustrated by cluster centers shown in Figure 2. We will see in Section 5.1 that such extreme situations do arise in real world data making it imperative that the proper distance metric be used for clustering distributional signatures.

5.1 Mobility Usage Signatures

Given the ubiquity of cellphones and their constant use, cell phone calling patterns have been used extensively to study the location and movements of large numbers of people for urban planning, event management and traffic monitoring. In addition, aggregated cell phone usage data find
application for traditional network management and monitoring tasks. Isaacman et al [8] show by using anonymized, aggregated statistics based on hundreds of thousands of cell phone calls in New York and Los Angeles, that the commuting patterns, travel distances and times are very different for New Yorkers and Los Angelinos. Similarly, by clustering anonymized aggregates of cell phone calls in a small town, Becker et al [2] identify interesting clusters of users with distinct behavior such as commuters and school going children.

In general, we are interested in analyzing mobile phone usage signatures in order to find common patterns, detect outliers, and identify trouble spots in the network. We describe below our experiments with the use of EMD for clustering mobile usage signatures. There exist many more applications of EMD to signatures, like clustering network traffic patterns and detecting changes in traffic distributions, but for the purpose of this paper, we confine ourselves to clustering usage signatures.

5.2 Data Description

The data consisted of over 400,000 unique handsets that generated around 15 million voice calls and 26 million text messages collected over a 60 day period from 35 cell towers located within a radius of 5 miles.

Given the sensitivity of the data, we took several steps to ensure the privacy of individuals. First, only anonymous records were used in this study. The data was collected and anonymized by a party not involved in the data analysis. We worked only on usage aggregates by hour and day of week. Second, all our results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of individuals.

For each user, we aggregated the calls and text messages separately over the 24 × 7 grid of “hour of day” vs. “day of the week” resulting in 168 bins. We had two such grids, one for voice and one for text message activity. The voice minutes and text message counts were made comparable by normalizing them so that they have the same global mean, resulting in one SMS corresponding to approximately 1.5 minutes of a voice call. The two usage vectors for each handset were concatenated to form one vector of 2 × 24 × 7 = 336 elements. This vector was further normalized to have a total sum of one. This then, was our signature, the distribution of cell phone usage across 336 bins. Examples of such distributions were presented in Figure 1, visualized as heat maps of cell phone usage across 336 bins. The resulting signatures are consistent with the within-cluster variance, and sparsity of the resulting distribution can be used to compute EMD for new data. This operation is computationally far less expensive than computing a full pairwise distance matrix. This is useful for several purposes. We can compare the resulting distances with the within-cluster distances to find outliers with respect to the existing data and clusters. In addition, we can use this information to assign new data into clusters.

Once the clusters have been formed, their signature distribution can be used to compute EMD for new data. This operation is computationally far less expensive than computing a full pairwise distance matrix. This is useful for several purposes. We can compare the resulting distances with the within-cluster distances to find outliers with respect to the existing data and clusters. In addition, we can use this information to assign new data into clusters.

5.3 Clustering the Signatures

The size of our data set consists of over 410,000 signatures which poses a challenge for clustering algorithms. Distance-based methods typically require a distance matrix consisting of \( (n^2 - n) / 2 \) entries, which in our case implies more than 84 billion distances. Although our proposed EMD approximation algorithm reduces the computational cost to linear time, the estimated duration for the computation of the 8.4 · 10^{16} distances was 6 days, therefore we decided to use a random sample of 50,000 signatures to perform the clustering. We can then use EMD to compute distances from cluster centers to the entire data set, allowing for both outlier detection and cluster assignment. The computation of EMD distances to each of the cluster centers for the entire large set is highly feasible and is completed in a matter of minutes.

A major advantage of using our novel approximation to EMD is that it allows us to define a topology on the bins, and introduce the notion of nearness of bins. Since the boundaries of bins by hour are sense arbitrary, we would consider neighboring bins in time as ‘close’. For example, we would consider users exhibiting activities at 12:45pm and 1:15pm more similar than those active at 7am and 11pm. Also typically human activities follow a daily cycle, so that there is a similarity between different days of the week at the same time. Finally, we have two types of activities – voice and SMS – which are distinct, so we model them as being at a constant distance from each other over the space of time. EMD allows us to specify such complex relationships between the bins.

To model the “time of day” and “day of week” relationship, we use a torus as the underlying geometry for the time and day of week space since the torus is defined as the product of two circles. By using a uniform helix on the torus we have a structure that wraps around at the end from one day to another and also from the end of the week to the beginning of the week. Finally we add the activity type as a product with \( \mathbb{R} \).

Figure 6 shows the result of a hierarchical clustering based on the EMD. The clustering was run on a random sample of 50,000 cases from the entire dataset, including a large portion of low usage signatures. We plot the aggregate signatures based on all members in each cluster. The cluster number is at the top left corner of each plot. The blue bar on top of each signature represents the relative size of the cluster. Since we are interested in detection of general patterns, the exact number of clusters is not critical for our application. We selected the number of clusters taking into account within-cluster variance, and sparsity of the resulting clusters.

The EMD clusters show distinct patterns. Cluster 1 corresponds to a mixed usage of both voice and SMS, predominantly during working hours. In contrast, Cluster 2 represents mostly SMS usage on evenings and into the weekend. Similar structure but limited to voice calls can be seen in Cluster 4. Clusters 3-6 separate signatures with almost no SMS activity whereas Clusters 7-9 have predominantly SMS usage. The clustering allows us to identify groups with similar usage patterns and their relative sizes, which can be used in many applications such as optimization of the operation of the cellular network, or on a local scale, to help in urban planning by identifying spatial groupings.

Once the clusters have been formed, their signature distribution can be used to compute EMD for new data. This operation is computationally far less expensive than computing a full pairwise distance matrix. This is useful for several purposes. We can compare the resulting distances with the within-cluster distances to find outliers with respect to the existing data and clusters. In addition, we can use this information to assign new data into clusters.

Figure 7 shows cluster signatures based on the entire dataset. We assigned each of the 411,550 signatures into clusters by taking the cluster with the closest center according to EMD. We see only a slight shift of the cluster size distribution between clusters compared to the clustering result on the subset of 50,000. The resulting signatures are consistent with the cluster centers, with some patterns standing out even
more sharply. Weekends are more strongly represented in Clusters 4 and 8, and Cluster 7 is more clearly focused on SMS during working hours. This allows us to extend our identification of patterns to the entire dataset.

In the same way that we used cluster signatures for comparison, any set of target distributions could be used, allowing supervised clustering applications on large scale as well. In our example, we used a fixed time frame, but if new data are obtained periodically, it allows for comparison of cluster evolution over time.

Alternatively, it is also possible to compare the evolution of signatures in linear time using our proposed EMD approximation by measuring the distance between distributions of subsequent signature in time. There are numerous application for EMD based analysis in change detection and spatio-temporal clustering.

6. CONCLUSION

In this paper, we proposed the use of the Earth Mover Distance for clustering distributions and signatures. EMD is inherently meaningful since it takes into account the underlying topology of the distributions unlike clustering methods like k-means that simply use the Euclidean metric defined over the point space. It allows for cross-bin comparisons creating a similarity measure that matches the perceptual notion of nearness.

EMD calculation usually require $O(n^3 \log n)$, but we proposed a novel generalization of a well-known wavelet approximation that enables us to apply EMD to massive data for general topologies including distributional signatures.

We applied EMD based clustering to a massive real world mobile phone usage data set of over 400,000 users by computing full pairwise distances, clustering of a sample subset of 50,000 and extending the results to the entire population. EMD provides meaningful results in every part of the usage distribution while k-means fails in clustering sparsely populated signatures.

6.1 Future Research

An interesting problem involves clustering multivariate time series. Existing techniques represent the time series as signatures and cluster them as was done in this paper, or impose artificial windows that capture slices of the time series and treat them as multi-dimensional vectors that are clustered in a routine way. While we ignore the temporal aspect in signatures, EMD can be used to cluster multivariate time series by defining similarity on the time space. Because of its notion of ‘nearness’ in the time domain it recognizes small shifts in the time series as being similar enough to belong to the same cluster.

Similarly, we can apply EMD to cluster sequences of discrete events like Twitter activity for a given topic.
Figure 7: Aggregate signatures of clusters for all 411,550 data points. Cases are assigned to clusters according to their proximity to cluster centers derived using EMD on a random sample of 50,000 signatures.

The approximation transforms the calculation of EMD into a computation of the $L_1$ norm between vectors in a sparse high-dimensional space. This permits exploiting the properties of the $L_1$ norm to further optimize applications, such as more efficient clustering algorithms without the need to compute the full pairwise distance matrix.

7. REFERENCES


