GRAPH-BASED SMOOTHING OF CLASS DATA WITH APPLICATIONS IN MUSICAL KEY FINDING

Olli Yli-Harja, Ilya Shmulevich

Signal Processing Laboratory
Tampere University of Technology
Tampere, Finland

Kjell Lemström

University of Helsinki
Department of Computer Science
Helsinki, Finland

ABSTRACT

We consider the problem of smoothing or estimating data which does not possess any numerical ordering properties. We propose a graph-based estimation method that does not rely on linear ordering of the data. The method is similar to L_p-norm estimates, such as the median and mean operations, and can be used to correct misclassifications. As an application, we consider the problem of determining the localized tonal context in a musical composition.

1. INTRODUCTION

Nonlinear digital filters have been successfully used for solving many types of problems in signal and image processing, especially in situations where linear filters are inappropriate. A large subclass of nonlinear filters that has been the focus of much attention is the class of filters based on order statistics [1]. This class includes median-related filters such as rank-order filters, order statistic filters, weighted median filters, stack filters, and morphological filters. Good overviews of these topics can be found in [2] and [3]. One property common to all such filtering schemes is that they inherently depend on some ordering of the data to be filtered. That is, a numerical ordering is intrinsic to the amplitude of one-dimensional signals or pixel values of images. However, some types of data may not possess any ordering property in that its constituents arise from what we will refer to generally as "classes." Such data will be referred to as "class data." An example of class data is an image in which every pixel is classified in one of a finite number of classes and is visually represented by some color. Such images can arise, for example, from classification of hyperspectral sensor data. The color itself is only symbolic in that it is used to represent the class (i.e. grass can be green, water can be blue, etc.). Naturally, the notion of "outlier" necessarily becomes somewhat vague as we have no recourse to any numerical information. Thus, outliers can be thought of simply as misclassifications.

Nevertheless, we may wish to process a sequence of class data with the goal of removing such outliers or misclassifications by utilizing the information in their neighborhood. After all, it is natural to expect that in many situations, neighboring pixels or signal values contain information about a central pixel or value. For example, if one pixel of grass is surrounded only by water pixels, it is possibly a misclassification and can be corrected. This serves as part of the motivation of this work. Specifically, we propose a new method of smoothing or estimating class data using a graph-based L_p-norm estimate. Furthermore, our method can incorporate distances between classes; that is, some classes may be more likely to be found in close proximity to each other than others. For example, grass may be more likely to be found close to water than roofs of buildings.

As an application of the proposed method, we consider the problem of determining the localized tonal context in a musical composition. The goal here is to trace varying tonal orientations as well as modulations in a reliable fashion. Our method is essentially based on a key finding algorithm developed by Krumhansl [4] which is applied in a sliding window fashion. Unfortunately, in practice, there is quite a bit of variation in certain regions of the sequence of key assignments. This is due to the algorithm's sensitivity to the distribution of pitches within the window and is a manifestation of the well known uncertainty principle. If the width of the window is made too large, the artifacts are suppressed, but the detection accuracy is sacrificed. On the other hand, smaller windows give rise to impulses and oscillations in the sequence of key assignments, while accuracy of detecting modulations may be improved. Therefore, the preferable action is to smooth out the local oscillations and to remove impulses by utilizing neighboring key assignments.

As a solution to this problem, various nonlinear fil-

ters, such as the recursive median filter have been employed [5]. As explained above, the difficulty with using such filters is due to the class quality of the input data. That is, there is no natural ordering of the tonal contexts. Addressing this problem, we apply the proposed graph-based method.

2. DEFINITIONS

Consider a complete undirected weighted graph $G(V, E)$ with vertex set V, edge set E and a weight function $w : V \times V \to \mathbb{R}$. Let us suppose that $w(v, v) = 0$ for all $v \in V$. In the case of real numbers, it is well known that the median of $(X_1, X_2, \ldots, X_n)$, $X_i \in \mathbb{R}$, is the value β minimizing

$$\sum_{i=1}^{n} |X_i - \beta|$$

So,

$$\text{med}\{X_1, X_2, \ldots, X_n\} = \arg \min_{\beta \in \{X_1, \cdots, X_n\}} \sum_{i=1}^{n} |X_i - \beta| \tag{1}$$

Similarly,

$$\text{mean}\{X_1, X_2, \ldots, X_n\} = \arg \min_{\beta} \sum_{i=1}^{n} (X_i - \beta)^2 \tag{2}$$

where β does not necessarily belong to $\{X_1, X_2, \ldots, X_n\}$. Suppose now that we have some set of samples $A = \{V_1, V_2, \ldots, V_n\}$, $V_i \in V$ of graph G. In a similar manner to (1) and (2), we can define

$$\text{graph-}p(A) = \arg \min_{\beta \in A} \sum_{i=1}^{n} w(V_i, \beta)^p \tag{3}$$

to be the graph-based L_p-norm estimate. The values of $p = 1$ and 2 correspond to graph-based median and mean, respectively. Note that the estimate is necessarily one of the vertices under consideration. Also, as the following example illustrates, vertices may be repeated; that is, it is possible that $V_i = V_j$ for $1 \leq i < j \leq n$.

Example 1 *Consider the graph shown in Figure 1. Suppose that the contents of our window are*

$$A = \{v_1, v_1, v_3, v_5, v_2\}$$

Let us compute graph-1(A). The sum of the weights from v_1 *to* v_1, v_3, v_5, *and* v_2 *is* 9 *(recall that* $w(v_1, v_1) = 0$*). The sum of the weights from* v_3 *to* v_1, v_1, v_5, *and* v_2 *is* 16. *The corresponding sums of the weights from* v_5 *and* v_2 *are* 17 *and* 19 *respectively. Therefore, since the sum of the weights is smallest from* v_1, *graph-1(A)* $= v_1$.

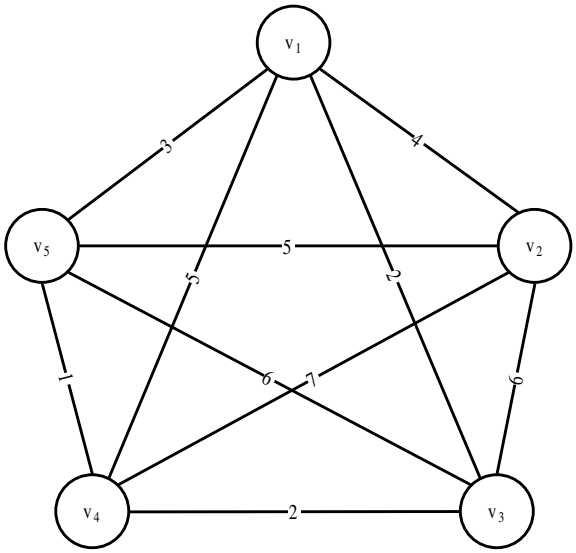

Figure 1: An example of a weighted graph. Weights are shown on edges.

Similarly to the running mean and median filters, we can define a sliding window filtering operation based on (3) as

$$Y_i = \text{graph-}p(X_{i-m}, \ldots, X_i, \ldots, X_{i+m}) \tag{4}$$

where $\{X_k\}$ is the sequence of input class data and $\{Y_k\}$ is the sequence of output class data, with $2m+1$ being the window width.

Finally, in certain situations, we may not have a complete graph at our disposal. In that case, we can construct a complete graph as follows. Suppose $G(V, E)$ is not a complete graph. For any pair of vertices $v_1, v_2 \in V$ such that $(v_1, v_2) \notin E$, we compute the shortest distance $d(v_1, v_2)$ from v_1 to v_2 and create an edge $e = (v_1, v_2)$ with weight $w(v_1, v_2) = d(v_1, v_2)$. If v_1 and v_2 are not connected in G, then $w(v_1, v_2) = \infty$.

3. APPLICATION TO KEY FINDING

As an application of the proposed approach, we consider the problem of determining the localized tonal context in a musical composition. The need for such an algorithm arises from a system for machine recognition of music patterns proposed in [6]. An important component of this system consists of determining a pitch error between a *target* (query) pattern and a *scanned* pattern from a music database. The pitch error consists of two parts: an objective or absolute component and a perceptual component. We focus on the latter and briefly review the necessary background.

3.1. A Key Finding Algorithm for Music Pattern Recognition

Performing classification based solely on the objective pitch error would not take into account the fact that intervals of equal size are not perceived as being equal when the tones are heard in tonal contexts [7]. Since the ultimate goal is to recognize a target pattern memorized (possibly incorrectly) by a human being, it is important to consider certain principles of melody memorization and recall. For example, findings showed that "less stable elements tended to be poorly remembered and frequently confused with more stable elements." Also, when an unstable element was introduced into a tonal sequence, "... the unstable element was itself poorly remembered" [4]. So, the occurrence of an unstable interval within a given tonal context (e.g., a melody ending in the tones C C♯ in the C major context) should be penalized more than a stable interval (e.g., B C in the C major context) since the unstable interval is less likely to have been memorized by the human user. These perceptual phenomena must be quantified for them to be useful in the classification of musical patterns. Such a quantification is provided by the *relatedness ratings* found by Krumhansl [4]. Essentially, a relatedness rating between tone q_1 and tone q_2 ($q_1 \neq q_2$) is a measure of how well q_2 follows q_1 in a given tonal context. The relatedness rating is a real number between 1 and 7 and is determined by experiments with human listeners. Results are provided for both major and minor contexts. So, a relatedness rating between two different tones in any of 24 possible tonal contexts can be found due to invariance under transposition.

To this end, suppose we are scanning a sequence of n notes to which we compare a target pattern consisting of n notes. For the moment, assuming knowledge of the tonal context of the scanned pattern, we define its vector of relatedness ratings $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{n-1}]$ as well as $\beta = [\beta_1, \beta_2, \cdots, \beta_{n-1}]$, the vector of relatedness ratings for the target pattern in the same tonal context. Each α_i and β_i is the relatedness rating between pitches q_i and q_{i+1} in the given tonal context for the scanned and target patterns respectively. Having defined the vectors of relatedness ratings for the scanned and target patterns, we can define the perceptual pitch error to be $e_p = \|\alpha - \beta\|_1$.

We have assumed that in the computation of the perceptual pitch error, we had knowledge of the tonal context of the scanned pattern. Thus, the need arises for a localized key finding algorithm which will present us with a most likely tonal context for a given musical pattern which will be subsequently used for the relatedness rating vectors. Such an algorithm was developed by Krumhansl [4] and is essentially based on the fact that "most stable pitch classes should occur most often" [8].

The algorithm produces a 24-element vector of correlations, $\mathbf{r} = [r_1, \cdots, r_{24}]$, the first twelve for major contexts and the others for minor contexts. The highest correlation, $r_{\max}$, is the one that corresponds to the most likely tonal context of the musical pattern being scanned. Suppose a musical composition (or set of compositions) that we wish to scan for the purpose of recognizing the target pattern consists of m notes and the target pattern itself consists of n notes (typically, $m \gg n$). In our algorithm, we slide a window of length n across the sequence of m notes and for each window position, the key-finding algorithm outputs a key assignment. Thus, we have a sequence $\mathbf{t} = [t_1, t_2, \cdots, t_{m-n+1}]$ of key assignments such that $t_i = \arg\max(\mathbf{r}_i)$. Figure 2 shows a typical sequence of key assignments.

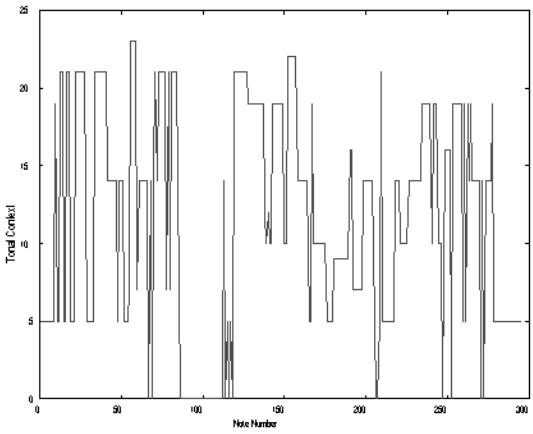

Figure 2: Typical sequence of key assignments. Since there is no natural ordering of tonal contexts, they are arbitrarily ordered for the purpose of visualization.

Unfortunately, in practice, there is quite a bit of variation in certain regions of the sequence of key assignments. Common artifacts are impulses and oscillations between modulations (edges). This is due to the algorithm's sensitivity to the distribution of pitches within the window. These small oscillations and impulses are undesirable, not only because they do not reflect our notions of modulations and localized tonal context, but primarily because they affect the relatedness rating vectors, which inherently depend on the tonal context produced by the key-finding algorithm.

Since the values of the assigned key sequence often appears arbitrary in the regions of oscillation, the perceptual pitch error is distorted in these regions. Therefore, the preferable action is to smooth out those local oscillations. As a solution to this problem, various nonlinear filters, such as the recursive median filter, have been employed [5]. As mentioned above, the problem with using such methods is that there is no natural total ordering of the tonal contexts. Thus, we see the need for an approach which doesn't depend on any ordering of the data.

3.2. Multidimensional Scaling and Graph-based Smoothing

At this point, the application of the proposed graph-based smoothing method seems straightforward. Each of the 24 tonal contexts (12 major and 12 minor) is represented by a vertex on a graph. However, the weights of the edges must represent interkey distances. In [9], correlations between so-called key profiles were used as a measure of interkey distances. A high correlation corresponded to a high degree of similarity between two keys while a low or negative correlation corresponded to a low degree of similarity. Then, the correlations were used to produce a spatial representation of distances between keys by using multidimensional scaling **??**. These resulting key distances provide a quantitative measure of similarity between all 24 tonal contexts.

The method transforms similarity values into a spatial representation of points in Euclidean space. The set of similarity values input to the multidimensional scaling program describes the degree of similarity or relatedness between all possible pairs of objects. The output of the multidimensional scaling program is a set of Euclidean coordinates for each of the tonal contexts. To objects which have high similarity values correspond to points which are close to each other in the coordinates space. The scaling program also provides a measure, called *stress*, that specifies how well the resulting configuration fits the similarity values. This stress value determines the number of necessary dimensions; as the number of dimensions increases, the stress value decreases. In this case, a four dimensional solution was found [4]. Two dimensions account for the circle of fifths while the other two account for parallel and relative major-minor relationships. Thus, the Euclidean distances in the spatial configuration in four dimensions represent interkey distances.

Now, we can set the weights of the edges to be distances from the multidimensional scaling solution. For example, the coordinate of C major is

$$[0.567, -0.633, -0.208, 0.480]$$

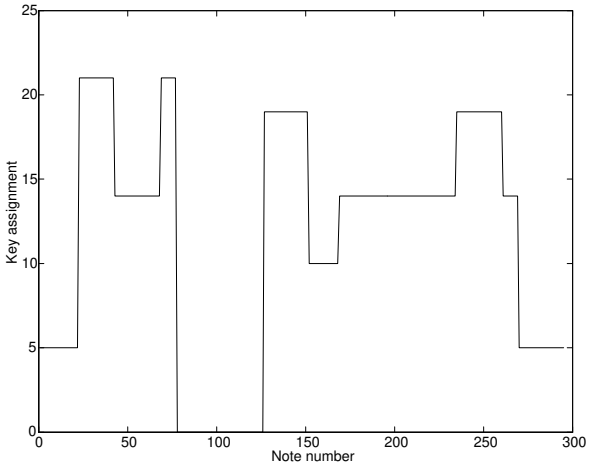

Figure 3: Graph-based L_1-norm estimates of key assignments

and the coordinate of A minor is

$$[0.206, -0.781, -0.580, 0.119]$$

Then, the Euclidean distance between these two keys is 0.6488, which is equal to the weight of the edge between those two vertices. Consider the following example.

Example 2 *Suppose that our window contains the following five key assignments: [C major; C major; C# major; C major, A minor]. We estimate the key assignment using the graph-based L_1-norm estimate, that is, graph-1$(\cdot)$. For each of the five keys, we compute and sum the distances to the other four keys. In order, they are*

$$[2.4491, 2.4491, 7.0923, 2.4491, 3.6377]$$

Then, we pick the key which had the minimum total distance to the rest of the vertices. In this case, the estimated key is C major.

The weights of the edges, or equivalently, interkey distances actually serve as additional factors contributing to the robustness of the estimation method. The estimator effectively takes into account the proximity of different classes. A class that has a high edge weights (distances) to the rest of the classes under consideration has a much lower chance of being selected than its neighbors. Figure 3 shows the application of this method in a sliding window fashion, as in equation (4), to the sequence of key-assignments shown in Figure 2. The window width in this case was equal to 37.

4. CONCLUSIONS

We have proposed a new method for smoothing or estimating class data, which is data that does not possess any intrinsic ordering. This method can be used to correct misclassifications and was demonstrated by applying it to a problem of determining the localized tonal context in a musical composition. One of the advantages of this method is that it allows one to incorporate a measure of closeness or proximity between classes.

As part of future work, several generalizations can be studied. For example, the graph can be generalized to be a directed graph and therefore, distances need not be symmetric. Also, the method should be applied to classification problems, such as to images of hyperspectral sensor data.

5. REFERENCES

[1] H. A. David, *Order Statistics*, Wiley, 1981.

[2] J. Astola, P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering*, CRC Press, 1997.

[3] I. Pitas, A. N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*, Kluwer Academic Publishers, 1990.

[4] C. L. Krumhansl, *Cognitive Foundations of Musical Pitch*, New York: Oxford University Press, 1990.

[5] I. Shmulevich, E.J. Coyle, "The Use of Recursive Median Filters for Establishing the Tonal Context in Music," *Proceedings of the 1997 IEEE Workshop on Nonlinear Signal and Image Processing*, Mackinac Island, MI, 1997.

[6] E.J. Coyle and I. Shmulevich, "A System for Machine Recognition of Music Patterns," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1998.

[7] C. L. Krumhansl, R. N. Shepard, "Quantification of the hierarchy of tonal functions within a diatonic context," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 5, pp. 579-594, 1979.

[8] A. H. Takeuchi, "Maximum key-profile correlation (MKC) as a measure of tonal structure in music," *Perception & Psychophysics*, vol. 56, pp. 335-346, 1994.

[9] C. L. Krumhansl, E. J. Kessler, "Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys," *Psychological Review*, vol. 89, pp. 334-368, 1982.

[10] J. B Kruskal, M. Wish, *Multidimensional Scaling*, Sage Publications, 1978.