Ontology-Based Mining of Brainwaves: A Sequence Similarity Technique for Mapping Alternative Features in Event-Related Potentials (ERP) Data

Haishan Liu[1], Gwen Frishkoff[2,3], Robert Frank[2], and Dejing Dou[1]

[1] Computer and Information Science Department, University of Oregon, University of Oregon, Eugene, OR 97403
[2] NeuroInformatics Center, University of Oregon, University of Oregon, Eugene, OR 97403
[3] Department of Psychology & Neuroscience Institute, Georgia State University, Atlanta, GA 30303

Abstract. In this paper, we present a method for identifying correspondences, or mappings, between alternative features of brainwave activity in event-related potentials (ERP) data. The goal is to simulate mapping across results from heterogeneous methods that might be used in different neuroscience research labs. The input to the mapping consists of two ERP datasets whose spatiotemporal characteristics are captured by alternative sets of features, that is, summary spatial and temporal measures capturing distinct neural patterns that are linked to concepts in a set of ERP ontologies, called NEMO (Neural ElectroMagnetic Ontologies) [3, 6]. The feature value vector of each summary metric is transformed into a point-sequence curve, and clustering is performed to extract similar subsequences (clusters) representing the neural patterns that can then be aligned across datasets. Finally, the similarity between measures is derived by calculating the similarity between corresponding point-sequence curves. Experiment results showed that the proposed approach is robust and has achieved significant improvement on precision than previous algorithms.

Keywords: Schema Matching, Sequence Similarity Search, ERP Data.

1 Introduction

Over the last two decades, neuroscience has witnessed remarkable advances in the development of methods for research on human brain function, including high-density electroencephalography (EEG) and event-related potentials (ERP). The ERP ("brainwave") method is a direct measure of neuronal activity.. ERP methods have yielded a large number of patterns that are associated with various behavioral and cognitive functions [12, 13]. Remarkably, however, there are few quantitative comparisons ("meta-analyses") of ERP data from different studies.. The inability to compare results across experiments has made it difficult to achieve a high-level synthesis and understanding of the vast majority of ERP results.

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 43–54, 2010.

To address this problem, we have been working to design a system, called Neural ElectroMagnetic Ontologies, or NEMO [3, 6], for data sharing and integration of results across different ERP analysis methods, experiment paradigms, and research sites with the help of formal ontologies. In the present paper, we extend this prior work by introducing a method for identifying correspondences, or mappings, between alternative sets of ERP spatial and temporal measures. These alternative measures reflect different ways that ERP pattern features can be summarized. For example, one research group might use a measure of peak latency (time of maximum amplitude) to summarize the timecourse of the "P100" pattern in a visual object processing experiment [14, 15], while another group might use measures of pattern onset and offset to characterize the same data. Given that different analysis methods may yield distinct and complementary insights, it is likely that this "embarrassment of riches" in ERP analysis will persist. The challenge then becomes how to develop an automatic way to find valid correspondences between features of ERP datasets that are derived from different analysis methods.

To this end, we create simulated ERP data using a tool that we develop, called NEMOautolabel[1]. We extract alternative measures of latency and scalp topography (see Appendix in [5] for example) to simulate heterogeneities that arise when distinct measure generation techniques are adopted by two different research groups. Our goal is then to discover mappings between the alternative measures. This is essentially a schema mapping (or matching, we use them interchangeably in the present paper) problem as the alternative sets of measures are served as features in different ERP datasets. Due to the nature of the ERP data, we face several unique challenges:

1. Useful schema information is limited, since the data under study is semi-structured.
2. Language-based or linguistic schema-level matcher that makes use of name and text similarity is not suitable, since alternative features of ERP datasets often use totally different names (see experiments in Section 4 for example).
3. Values of alternative measures are numerical. Conventional instance-level matcher that handles mapping between numerical elements based on extraction of statistical characterization, such as range, mean and standard deviation, are not appropriate, since they are too rough to capture patterns that are crucial in determining the correspondence.

To address these challenges, we propose a novel method that explores sequence similarity search techniques and the NEMO ontologies, resulting in a framework for *ontology-based mining* of ERP data. Ontology-based mining has recently emerged as a new area of data mining, in which ontologies are used as formal domain knowledge to guide the data mining process in order to enhance performance and to represent the data mining result. Our method starts by transforming the vector of values of each measure into a *point-sequence curve*, and then evaluates similarities of the curves across datasets to determine the appropriate mapping across measures. The key problem then becomes to align subsequences of values in a principled way, thus enabling valid comparisons among instances of spatial and temporal measures across datasets. If the correspondence between two measures is not known a priori (as assumed in the present study), and if the values for these two measures are plotted

[1] http://nemo.nic.uoregon.edu/wiki/NEMO_Analysis_Toolkit

against the arbitrary instance numbers associated with the two datasets, the resulting graph will show no clear pattern and thus no correspondence between alternative measures (see Fig. 1, left frame). Our solution is to introduce structure into these (initially random) point-sequence curves by applying clustering to extract similar subsequences, which are further labeled using terms defined in the NEMO ontologies. These subsequences can then be aligned across the datasets, and correspondences between measures established using standard techniques for time-sequence similarity search (see Fig. 1, right frame). This approach exploits prior (domain) knowledge of the patterns that are commonly seen in ERP experiments of a particular type (e.g., visual perception) while asserting no prior knowledge about the measures.

The rest of this paper is organized as follows: In Section 2 we give a brief overview of prior work on schema matching with a focus on instance-level approaches, and time-sequence similarity search. In Section 3 we present the simulated ERP data design and methods for point-sequence matching. In Section 4, we present the ERP mapping results. Finally, in Section 5, we consider the assumptions and constraints of these methods and discuss future research directions, highlighting the contributions of this work to future research on schema matching and meta-analysis of ERP data.

2 Related Works and Background

2.1 Schema Matching

Our study of mapping alternative measure sets is closely related to the schema matching problem. A schema matching algorithm may use multiple matching methods or *matchers*. It generally falls into one of two categories based on if it considers instance data or only schema information. Our ontology-based mining approach should be considered as one kind of instance-level method. According to the type of instance value, various instance-based approaches have been developed in previous research. For example:

- For textual attributes, a linguistic characterization based on information retrieval techniques can be applied [18].
- For nominal attributes, evaluation of the degree of overlap of instance values is a preferred approach. Larson *et al.* [10] and Sheth *et al.* [11] discussed how relationships and entity sets could be integrated primarily based on their domain relationships: EQUAL, CONTAINS, OVERLAP, etc. Similarity of partially overlapped instance set can be also calculated based on measures such as Hamming distance and Jaccard coefficient.
- For numeric attributes, typically one can use their values to compute statistics to characterize the attributes—e.g., 'SSN' and 'PhonNo' can be distinguishable since their data patterns, such as value distributions, and averages, are different [18].

Hybrid systems that combine several approaches to determine matching often achieve better performance. For example, SemInt [16, 17] is a comprehensive matching prototype exploiting up to 15 constraint-based and 5 content-based matching criteria. Instance data is used to enhance schema-level information by providing actual value

distributions, numerical averages, etc. SemInt determines a *match signature* for each attribute for either all or a selected subset of the supported criteria. Then neural networks or distance-based similarity measures over signatures can be used for determining an ordered list of match candidates.

The LSD (Learning Source Descriptions) system uses machine-learning techniques to match a new data source against a previously determined global schema [18]. It represents a composite match scheme with an automatic combination of match results. In addition to a name matcher they use several instance-level matchers (learners) that are trained during a preprocessing step. Given an initial user-supplied mapping from a data source to the global schema, the system trains multiple learners, thereby discovering characteristic instance patterns. These patterns and rules can then be applied to match other data sources to the global schema.

The iMAP [9] system can semi-automatically discover one-to-one and even complex mappings between relational database schemas. The goal is to reformulate the matching problem as a search in a match space, To perform the search effectively, iMAP uses multiple basic matchers, called searches, e.g., text, numeric, category, unit conversion, each of which addresses a particular subset of the match space.

An important limitation of the above instance-based matching methods is their inability to properly handle numerical instances in some certain domain application. They use statistical characterization extracted from the numerical instances, such as range, mean and standard deviation, to determine match. However such information is too rough to capture patterns in ERP data that are crucial in determining the correspondence. By contrast, our proposed sequence similarity search technique is specifically designed to handle attributes with numerical values for ERP data: a spatial distance measure is used to calculate the similarity between point-sequence curves representing the numerical attributes after subsequence reordering based on clustering, as described in Section 3.

2.2 Subsequence Similarity Search

We assume that similarity between point-sequence curves implies similarity between the metrics they represent. Therefore, we view the discovery of mappings between metric sets as a similarity search among two sets of point-sequence (time series) data.

Sequence similarity search has emerged as an active area of research. In general, methods for sequence similarity search belong to one of two categories [1]: 1) Whole Matching—the sequences to be compared have the same length (after interpolation or offset adjustment if necessary); and 2) Subsequence Matching—the query sequence is smaller; we look for a subsequence that best matches the query sequence.

The ERP metric mapping problem is a whole matching problem. Furthermore, we consider the cross spatial distance join [4] problem as a special case of whole matching. The spatial distance join is defined using two datasets, A and B, and a distance function *L*. For a given radius *r*, the spatial distance join computes the following set:

$$\{\langle a, b\rangle | a \in B, L(a, b) \leq r\} .$$

The term cross spatial join emphasizes that the two point sets A and B are distinct. The distance function *L* represents a similarity measure.

Performing the sequence similarity search task consists primarily of making the following choices: 1) a distance function *L*; 2) a method to generate cross pairs (*a*, *b*); and 3) a usage of approximations of objects as an index to the exact representation (i.e., to calculate *L*). Agrawal *et al*. [1] point out that the choice of *L* is clearly application-dependent. Although a wide spectrum of similarity measures has been proposed, a comprehensive survey by Keogh et al [7], which carried out extensive performance tests on different similarity measures, demonstrated that Euclidean distance outperformed other distance metrics. Therefore, we chose Euclidean distance as the distance function *L* in our study.

The problem of performing efficient spatial joins in relational database systems has been studied by Brinkho *et al*. [2]. They point out that spatial join is a kind of multiple-scan query where objects have to be accessed several times and therefore, execution time is generally not linear but superlinear in the number of objects. They propose to use the R-tree family to support efficient spatial queries and manage to achieve almost optimal I/O time.

For performance issues, indexing is also essential for similarity searches in sequence databases. Indexing is a technique that extracts k features from every sequence, maps them to a k-dimensional space, and then discovers how to store and search these points. This method can help alleviate the "curse of dimensionality" and to preserve spatial locality in disk pages for I/O optimization.

In the present study, we adopt a "naïve" approach that computes similarity on every cross-join pair of conjugate sequences. The cross join is performed by multiple sequential scans of the two datasets, and we do not perform indexing on the original sequences. The rationale is that scalability is not a major concern since the number of sequences (i.e., number of measures) in most ERP datasets is relatively small (<20).

3 Methods

We propose to view the feature value vector of each ERP summary metric as forming a point-sequence curve. The problem of matching discovery between metric sets can then be framed as a sequence similarity search task. To identify structured subsequences in each feature vector, we use clustering and label discovered clusters with respect to the simulated ERP patterns or "components" (e.g., *P100*, *N100*, *N3*, *MFN*, and *P300*. All of them are defined in the NEMO ontologies). By labeling the feature instances in this way, we can group them in each dataset based on their pattern labels and then align the instance groups across datasets accordingly. This step can be viewed as a subsequence reordering process. We then apply a sequence post-processing step to achieve better performance in the similarity search, leveraging the rich collection of sequence similarity search algorithms presently available. The final step is to evaluate the similarity of the structured point-sequence curves that now represent our two simulated ERP datasets as quantified by their respective measures. This evaluation is achieved by using the cross-spatial join to calculate the distance between all pairs of sequences from the two datasets. In this way, we can discover matching pairs of measures. Each of these steps is described in the following sections.

3.1 Simulated ERPs

The raw data for this study consist of 80 simulated event-related potentials (ERPs), where each ERP comprises simulated measurement data at 150 time samples and 129 channels (electrodes) for a particular subject (n=40) and experiment condition (n=2). The 40 simulated subjects are randomly divided into two datasets, *SG1* and *SG2*, each comprising 40 ERPs (20 subjects and 2 experimental conditions). Each ERP consists of a superposition of 5 latent spatiotemporal patterns that represent the scalp projections of distinct neuronal groups (dipoles). To create these patterns of neural activity, 9 dipoles are located and oriented within a 3-shell spherical model to simulate the topographies of 5 ERP components commonly seen in studies of visual word recognition. Each dipole is then assigned a 600 ms activation consistent with the temporal characteristics of its corresponding ERP. Simulated "scalp-surface" electrode locations are specified with a 129-channel montage, and a complex matrix of simulated noise is added to mimic known properties of human EEG. Because of volume conduction and the overlap of their temporal activity, the dipole activations induce a complex spatial and temporal superposition of the 5 modeled ERP patterns.

Spatiotemporal components are extracted from the two datasets, SG1 and SG2, using two techniques: temporal Principal Components Analysis (tPCA) and spatial Independent Components Analysis (sICA), two data decomposition techniques that are widely used in ERP research. Two alternative metric sets, *m1* and *m2*, are subsequently applied to the two tPCA and the two sICA derived datasets to quantify the spatiotemporal characteristics of the extracted patterns.

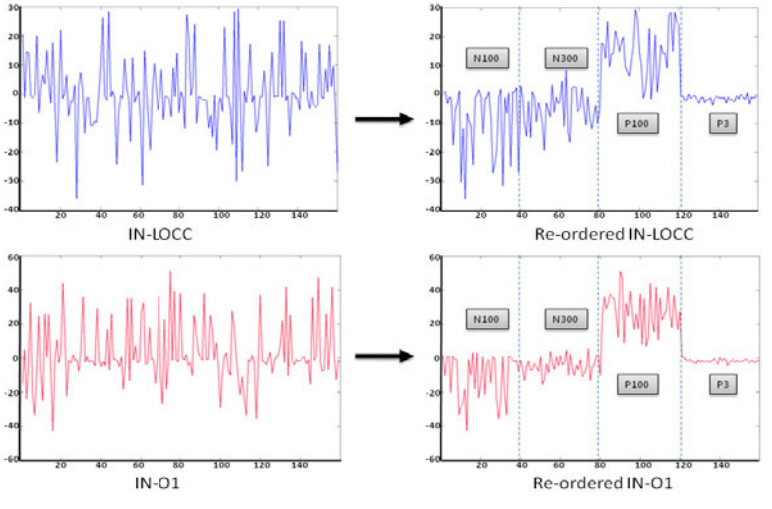

Fig. 1. (Left) *IN-LOCC* and *IN-O1* point-sequence curves prior to grouping and reordering. (Right) Labeled curves for metrics *IN-O1* and *IN-LOCC* after grouping/reordering.

3.2 Data Partitioning and Reordering

In the present study, we perform clustering on the spatial and temporal values of the two alternative sets of measures using Expectation Maximization (EM) algorithm. The resulting clusters represent candidate ERP patterns, characterized by the central tendencies of their cluster attributes (i.e., mean values for the spatial and temporal metrics). We label the resulting clusters with pattern labels defined in the NEMO ontologies (P100, N100, etc.) using rules specified by domain experts.

Following clustering and labeling, the pattern labels are used to align groups of instances across datasets, resulting in subsequence reordering. As illustrated in the right-hand graphs of Fig. 1, the point-sequence curves for metrics *IN-O1* and *IN-LOCC* (plotted using their original orderings prior to grouping/reordering on the left-hand side) are manifestly more similar after reordering subsequences in the two curves by aligning instances that belong to the same (or similar) patterns.

3.3 Sequence Post-processing

After alignment of the subsequences according to pattern labels defined in the NEMO ontology, we carry out three post-processing steps: (1) Normalization, i.e., scaling all the sequence values to unit range; (2) Smoothing, using a moving average method to reduce within cluster variance; and (3) Interpolation of curves, if the number of points in two point-sequence curves is different. Fig. 2 illustrates the results of normalization, smoothing and interpolation to the point-sequence curves of *IN-O1* and *IN-LOCC* in Fig. 1.

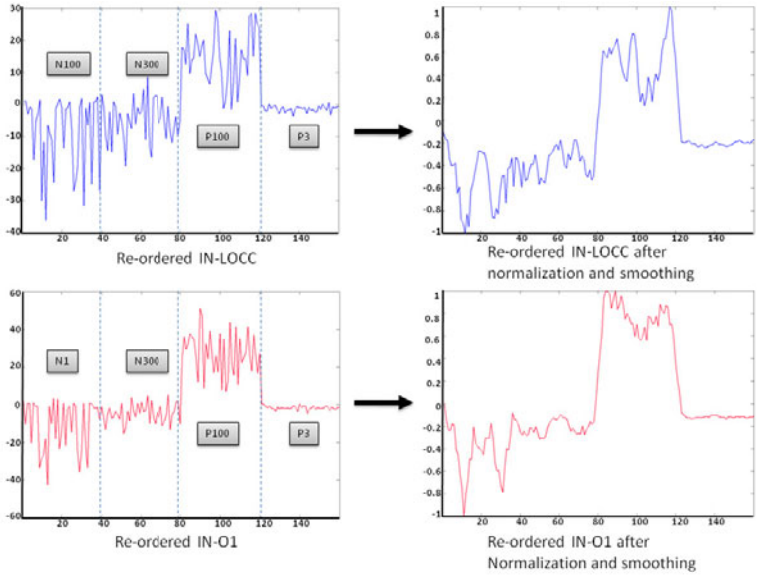

Fig. 2. After normalization, smoothing, and interpolation of point-sequence curves in Fig. 1

3.4 Sequence Similarity Search

The following heuristic assumptions are adopted in our sequence matching procedure.

First, we assume that the two datasets from which these alternative measures are extracted contain the same or similar ERP patterns. This assumption is critical, since it permits us to reorder the two point-sequence curves by aligning subsequences that are associated with the same ERP pattern labels.

Second, we assume that there exists a 1-to-1 mapping between pairs of metrics from the alternative sets of metrics. In other words, there must be no cells selected within the same column.

Table 1. Example for violation of the 1-1 mapping assumption and the solution

	IN-O1	**IN-O2**
IN-LOCC	4.08	***3.74***
IN-ROCC	4.01	***3.57***

(a)

	IN-O1	**IN-O2**
IN-LOCC	***4.08***	3.74
IN-ROCC	4.01	***3.57***

(b)

For example, Table 1(a) illustrates a scenario where the 1-to-1 mapping assumption is violated: the value in each cell is the Euclidean distance between two point-sequence curves denoted by the row and column header of the cell. If we select cells with minimum distance value in each row, we end up with two cells within the same column being selected, suggesting that both *IN-LOCC* and *IN-ROCC* are mapped to *IN-O2* in the present case. Table 1(b) illustrates the solution: cells are selected using the 1-to-1 mapping heuristic coupled with the global minimum heuristic (see below).

Finally, we assume a global minimum heuristic: we select those cells whose Euclidean distance values sum up to a minimum value.

Table 2. Solution to Table 1 using global minimum heuristic

	IN-O1	**IN-O2**
IN-LOCC	4.08	***3.74***
IN-ROCC	***4.01***	3.57

(a)

	IN-O1	**IN-O2**
IN-LOCC	***4.08***	3.74
IN-ROCC	4.01	***3.57***

(b)

For example, Table 2 shows two alternative cell selections that do not violate the 1-to-1 mapping heuristic. The global minimum heuristic requires us to favor 2(b) because $4.08 + 3.57 < 3.74 + 4.01$. The cell selections that achieve the global minimum suggest the most stable mapping result. The global minimum heuristic requires a non-greedy implementation that should take into consideration all possible selections. When the number of metrics is large, this implementation becomes more computationally challenging.

4 Results

The experiment is conducted on the simulated datasets described in Section 3.1. The test cases for the matching discovery experiment are derived as follows: each test case contains a source and target dataset that are derived respectively from one subject group (SG1 or SG2) characterized with one metric set (m1 or m2) and formulated under one decomposition method (sICA or tPCA), and from the other subject group with the alternative metric set and decomposition method. This yields 2 (subject groups) × 2 (metric sets) × 2 (decomposition method) = 8 test cases, each of which includes two different datasets, two alternative metric sets and two decomposition methods. In order to test the robustness of the proposed methods, we replicate the datasets for each test case into five copies with different random ordering of the instances, thus resulting in a total of 40 enriched test cases.

We test our method on each of these test cases. Table 3, for example, shows a distance table calculated by cross-spatial join of tPCA-derived data from SG1-m1 and SG2-m2. The highlighted cells indicate similarity pairs between two point-sequence curves representing two measures (row header and column header which meet at this cell) and are selected by using the 1-to-1 mapping and global minimum heuristics described in Section 3.4. A similarity pair represents a potential mapping discovered by our methods. For example, from this table we derive the following mappings: *IN-O1*↔*IN-LOCC*, *IN-O2*↔*IN-ROCC*, *IN-C3*↔*IN-LPAR*, etc. Note that the orders of the row and column header labels are such that the golden standard mapping falls along the diagonal cells. Therefore we can easily conclude that the precision of mapping in this test case is 9/13=69.2% since 4 out of 13 cells are shifted off from the diagonal.

Table 3. Cross-spatial join of data from SG1-m1 (tPCA) and SG2-m2 (tPCA)

	IN-O1	**IN-O2**	**IN-C3**	**IN-C4**	**IN-T7**	**IN-T8**	**IN-F7**	**IN-F8**	**IN-Fp1**	**IN-Fp2**	**IN-F3**	**IN-F4**	**TI-max2**
IN-LOCC	***2.76***	2.76	8.59	8.52	9.68	10.44	11.52	11.61	11.56	11.56	7.92	7.90	12.93
IN-ROCC	2.75	***2.75***	8.58	847.00	9.69	10.47	11.55	11.64	11.60	11.60	7.91	7.86	12.95
IN-LPAR	8.57	8.58	***4.13***	5.12	9.29	8.91	9.24	9.07	8.98	8.97	5.58	6.07	9.39
IN-RPAR	7.97	7.97	3.55	***4.38***	8.97	8.66	9.10	8.93	8.88	8.85	4.99	5.39	9.43
IN-LPTEM	9.32	9.34	8.54	9.23	5.00	***4.26***	5.62	5.34	5.73	5.72	7.37	7.88	11.42
IN-RPTEM	7.81	7.81	7.66	8.05	***4.18***	3.84	5.61	5.39	5.85	5.78	6.24	6.56	11.28
IN-LATEM	11.00	11.00	8.40	8.96	3.20	2.74	***2.30***	2.09	2.52	2.43	6.89	7.35	10.95
IN-RATEM	11.19	11.19	8.53	9.03	3.33	2.45	2.51	***2.08***	2.80	2.64	6.99	7.41	11.30
IN-LORB	9.58	9.58	6.00	6.48	4.23	4.50	3.58	3.63	3.35	***3.26***	4.36	4.83	10.31
IN-RORB	11.19	11.20	8.36	8.93	3.44	3.33	2.15	2.12	***2.21***	2.16	6.85	7.33	10.83
IN-LFRON	6.72	6.71	4.05	4.01	6.30	7.10	6.91	7.06	6.76	6.71	***2.74***	2.20	9.99
IN-RFRON	6.36	6.33	4.58	4.03	7.09	7.94	8.01	8.15	7.96	7.88	3.42	***3.06***	10.67
TI-max1	11.72	11.71	7.18	7.74	12.12	11.74	12.02	11.88	11.89	11.87	9.36	9.61	***8.58***

The performance of our methods among the 40 test cases is quite good. Table 4 summarizes the precision for each test case. The table consists of eight divisions, each of which illustrates the precision measures for the datasets generated by five samples of replication to one of the original eight test schemes with random instance ordering. Since the fact that the precision of mapping by making a random guess is almost zero and that the results demonstrate consistent performance on randomly ordered data, the

Table 4. Precision results for 40 test cases

(SG1, sICA, m1) *vs*. (SG2, sICA, m2)	
Input	Precision
Sample 1	13/13
Sample 2	13/13
Sample 3	13/13
Sample 4	13/13
Sample 5	13/13

(SG1, tPCA, m1) *vs*. (SG2, tPCA, m2)	
Input	Precision
Sample 1	9/13
Sample 2	9/13
Sample 3	9/13
Sample 4	9/13
Sample 5	9/13

(SG2, sICA m1) *vs*. (SG1, sICA, m2)	
Input	Precision
Sample 1	9/13
Sample 2	9/13
Sample 3	9/13
Sample 4	9/13
Sample 5	9/13

(SG2, tPCA, m1) *vs*. (SG1, tPCA m2)	
Input	Precision
Sample 1	9/13
Sample 2	9/13
Sample 3	9/13
Sample 4	9/13
Sample 5	9/13

(SG1, sICA, m1) *vs*. (SG2, tPCA, m2)	
Input	Precision
Sample 1	13/13
Sample 2	13/13
Sample 3	13/13
Sample 4	13/13
Sample 5	13/13

(SG1, tPCA, m1) *vs*. (SG2, sICA, m2)	
Input	Precision
Sample 1	5/13
Sample 2	5/13
Sample 3	5/13
Sample 4	5/13
Sample 5	5/13

(SG2, sICA, m1) *vs*. (SG1, tPCA, m2)	
Input	Precision
Sample 1	5/13
Sample 2	8/13
Sample 3	5/13
Sample 4	5/13
Sample 5	5/13

(SG2, tPCA, m1) *vs*. (SG1, sICA, m2)	
Input	Precision
Sample 1	7/13
Sample 2	7/13
Sample 3	7/13
Sample 4	7/13
Sample 5	7/13

precision of our method appears markedly robust. Combining the mapping results in the 40 test cases into an ensemble model by a majority vote of each individual mapping, we obtain the ensemble mapping result. The overall precision is 11/13=84.6%.

We compare the performance our algorithm with SemInt [16, 17] as the baseline. Since the data contains only numerical instances, SemInt extracts from each feature value vector 5 discriminators, namely, MIN, MAX, Average, Coefficient of variance, and Standard Deviation. The feature value vector is then projected to a *match signature* characterized by these discriminators. A neural network is trained based on datasets from the 40 test cases with one metric set and tested on the rest datasets with the alternative metric set to determine the match. The result shows 19.23% precision. As we point out in Section 1, the reason why our algorithm significantly outperforms SemInt is that we are able to systematically exploit prior knowledge about patterns in ERP data that is crucial to determine the matching.

5 Conclusion and Future Work

In this paper, we describe a method for identifying correspondences, or mappings, between alternative sets of ERP measures that might be used by different ERP research labs to characterize patterns of brain electrical activity. The contributions of this work include the following:

- Use of an ontology to assign meaningful labels to ERP patterns (clusters) and thereby impose structure that is used to align alternative metrics across datasets;
- Application of sequence similarity search in discovering mappings across alternative metrics;
- Extension of the instance-level approach in schema matching, especially to handle numerical values; and
- Articulation of a global minimum heuristic in selecting 'similarity pairs' from the distance table. This heuristic proved to be useful and empirically robust in our experiment.

Mappings between alternative spatial and temporal metrics can be used to link different representations of ERP data and thus to support representations of ERP results with the help of formal ontologies [6]. In this way, our work is closely related to schema/ontology matching, which has been an active research field for a number of years [8, 18]. In the course of developing and testing these methods, we have collected a corpus of real data from different experiments [5] and have observed a large number of different kinds of heterogeneities. The presence of these heterogeneities suggests that a method for identifying mappings between features or metrics across two datasets may have widespread applications for ontology-based integration, beyond the specific applications discussed in the present study.

Following we summarize some basic assumptions and limitations of the current study and then discuss some possible directions for future work.

The proposed method assumes some domain-specific knowledge, as well as certain features of the input data. First, the source and target datasets are assumed to contain the same or similar ERP patterns. If the two datasets contain dissimilar patterns, there will be few instances that can be aligned according to common pattern labels, resulting in a poor sequence similarity search result. Second, there is assumed to be a 1-to-1 mapping between alternative data metrics. This assumption may be violated in many real-world cases. For example, in ERP data, the temporal metrics TI-begin and TI-end together capture the same information as the metric TI-duration. Our method will need to be modified in the future to handle these more complex mappings.

Other challenges include the scalability of calculations for the global minimum in the distance table, which is essentially an NP-hard problem. It could be remedied by proper implementation such as dynamic programming, but remains computationally intractable when the number of metrics is very large. Future work will seek to find an appropriate approximation method that balances the interest in accuracy and scalability. In addition, the simulated ERP data used in the present study were carefully designed to mimic many, but not all, features of real ERP datasets. In particular, we minimized variability in latency and spatial distribution of patterns across the different ERPs so that the data decomposition and clustering of patterns would remain tractable and relatively straightforward to interpret. In future work, we plan to carry out more substantial tests on genuine ERP datasets, such as those that have been collected, analyzed, and stored in our NEMO ERP ontology database [20].

References

1. Agrawal, R., Faloutsos, C., Swami, A.N.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730, pp. 69–84. Springer, Heidelberg (1993)
2. Brinkhoff, T., Kriegel, H.-P., Seeger, B.: Efficient processing of spatial joins using r-trees. In: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 237–246. ACM, New York (1993)
3. Dou, D., Frishkoff, G., Rong, J., Frank, R., Malony, A., Tucker, D.: Development of NeuroElectroMagnetic Ontologies (NEMO): A Framework for Mining Brain Wave Ontologies. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2007), pp. 270–279 (2007)

4. Faloutsos, C., Seeger, B., Traina, A., Traina Jr., C.: Spatial join selectivity using power laws. In: SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 177–188. ACM, New York (2000)
5. Frishkoff, G.A., Frank, R.M., Rong, J., Dou, D., Dien, J., Halderman, L.K.: A Framework to Support Automated Classification and Labeling of Brain Electromagnetic Patterns. Computational Intelligence and Neuroscience (CIN), Special Issue, EEG/MEG Analysis and Signal Processing 2007 13 (2007)
6. Frishkoff, G., Le Pendu, P., Frank, R., Liuand, H., Dou, D.: Development of Neural Electromagnetic Ontologies (NEMO): Ontology-based Tools for Representation and Integration of Event-related Brain Potentials. In: Proceedings of the International Conference on Biomedical Ontology, ICBO 2009 (2009)
7. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Min. Knowl. Discov. 7(4), 349–371 (2003)
8. Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hubner, S.: Ontology-based integration of information: A survey of existing approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing, pp. 108–117 (2001)
9. Dhamankar, R., Lee, Y., Doan, A., Halevy, A.Y., Domingos, P.: iMAP: Discovering Complex Mappings between Database Schemas. In: Proceedings of the ACM Conference on Management of Data, pp. 383–394 (2004)
10. Larson, J., Navathe, S., Elmasri, R.: A theory of attributed equivalence in databases with application to schema integration. IEEE Transactions on Software Engineering 15(4), 449–463 (1989)
11. Sheth, A., Larson, J., Cornelio, A., Navathe, S.: A tool for integrating conceptual schemas and user views. In: Proc. 4th International Conference on Data Engineering (ICDE), Los Angeles, CA, US, pp. 176–183 (1988)
12. Gratton, G., Coles, M.G.H., Donchin, E.: A procedure for using multi-electrode information in the analysis of components of the event-related potential: Vector filter. Psychophysiology 26(2), 222–232 (1989)
13. Spencer, K.M., Dien, J., Donchin, E.: A componential analysis of the ERP elicited by novel events using a dense electrode array. Psychophysiology 36, 409–414 (1999)
14. Donchin, E., Heffley, E.: Multivariate analysis of event-related potential data: a tutorial review. In: Otto, D. (ed.) Multidisciplinary Perspectives in Event-Related Brain Potential Research, pp. 555–572. U.S. Government Printing Office, Washington (1978)
15. Picton, T.W., Bentin, S., Berg, P., et al.: Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. Psychophysiology 37(2), 127–152 (2000)
16. Li, W., Clifton, C.: Semantic integration in heterogeneous databases using neural networks. In: Proc. 20th Intl. Conf. on Very Large Data Bases, pp. 1–12 (1994)
17. Li, W., Clifton, C.: SemInt: a tool for identifying attribute correspondences in heterogeneous databases using neural network. Data Knowl. Eng. 33(1), 49–84 (2000)
18. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
19. Doan, A.H., Domingos, P., Halevy, A.: Reconciling schemas of disparate data sources: a machine-learning approach. In: Proc ACM SIGMOD Conf., pp. 509–520 (2001)
20. LePendu, P., Dou, D., Frishkoff, G., Rong, J.: Ontology Database: A New Method for Semantic Modeling and an Application to Brainwave Data. In: Ludäscher, B., Mamoulis, N. (eds.) SSDBM 2008. LNCS, vol. 5069, pp. 313–330. Springer, Heidelberg (2008)