Combining Expert Representations and Neural Networks for Visualization of Clinical Data

David M. Fram
Michael F. Johnston

Stephen I. Gallant
Channing H. Russell

Belmont Research Inc.
Cambridge, Massachusetts

1 Introduction

Neural network visualization and learning algorithms provide an attractive approach to displaying clinical data and predicting patient outcomes. However, the raw data in medical databases is typically not in a form that permits easy application of such algorithms. For example, data may consist of scattered and noisy individual measurements that are determined by available tests and instrumentation. In order to use this data for visualization or for prediction by human or machine, it may be necessary to combine several measurements over appropriate time periods.

One way to overcome this problem is by having a human expert create an *expert representation* consisting of a fixed-length feature vector. Each component of the vector is a relevant scalar, Boolean, or enumeration (e.g. "meal size") value that is useful for the problem at hand. Vector components are determined by a knowledgeable human expert, where each component might be derived from several different types of raw clinical data and/or data from several time periods. For example with the diabetes dataset used here, the components included the total insulin dose over the past three days, several blood glucose measurements, etc. Vector components can be partly redundant, because subsequent clustering and learning algorithms can weight components appropriately. This eases the task of the expert, because he or she can concentrate on features *sufficient* for prediction tasks, with little concern over *independence* of features.

Once an appropriate set of features has been determined and software produced to transform raw data to feature vectors, *all remaining processing is automatic.* The feature vector representation is well suited for building neural network models and for prediction, alarming and display based upon those models. Note that it may be helpful to tune the feature set based upon insight gained from models and their predictions.

For this paper we decided to use a diabetes dataset to experiment with an expert representation for cluster-based visualization. The next Section gives a quick overview of our work and some important details. An example is then presented, followed by concluding remarks.

2 Experiments

One of the authors (Johnston), an MD, constructed a set of features for characterizing the status of a patient on a particular day. We then used Kohonen's Topology Preserving Map algorithm [Kohonen 82a,b,88; see also Gallant 93] to cluster the data. Finally, we wrote code to display the set of clusters with respect to particular queries.

Feature Selection

The data set consisted of reports from 70 patients. From 8 days to 166 days of data were provided for each patient, consisting of blood glucose measurements, insulin type and dosage, meals, exercise, and hypoglycemic symptoms. The completeness of the data varied widely among patients.

One of our primary interests is the visualization and interpretation of time-oriented clinical data. From the raw data, we computed daily parameters representing the current day, the previous day, and averages over the previous three days. The feature vectors for input to the learning algorithm each represented one *patient-day* and contained the following fields for *each of the three time intervals*:

patient number	(ignored)
day	weekday/weekend
maximum blood glucose	mg/dl
minimum blood glucose	mg/dl
mean blood glucose	mg/dl
total insulin dose	units
regular insulin dose	units
NPH insulin dose	units
UltraLente insulin dose	units
breakfast	large/small/typical/ unknown
lunch	large/small/typical/ unknown
dinner	large/small/typical/ unknown
exercise	more/less/typical/ unknown
hypoglycemic symptoms	present/unknown
change in insulin dose from previous day	units

By this method, we constructed 3721 example vectors of 39 elements. Each field in the feature vectors was then normalized by subtracting the mean for the data set and dividing by the standard deviation.

Kohonen's Topology Preserving Maps

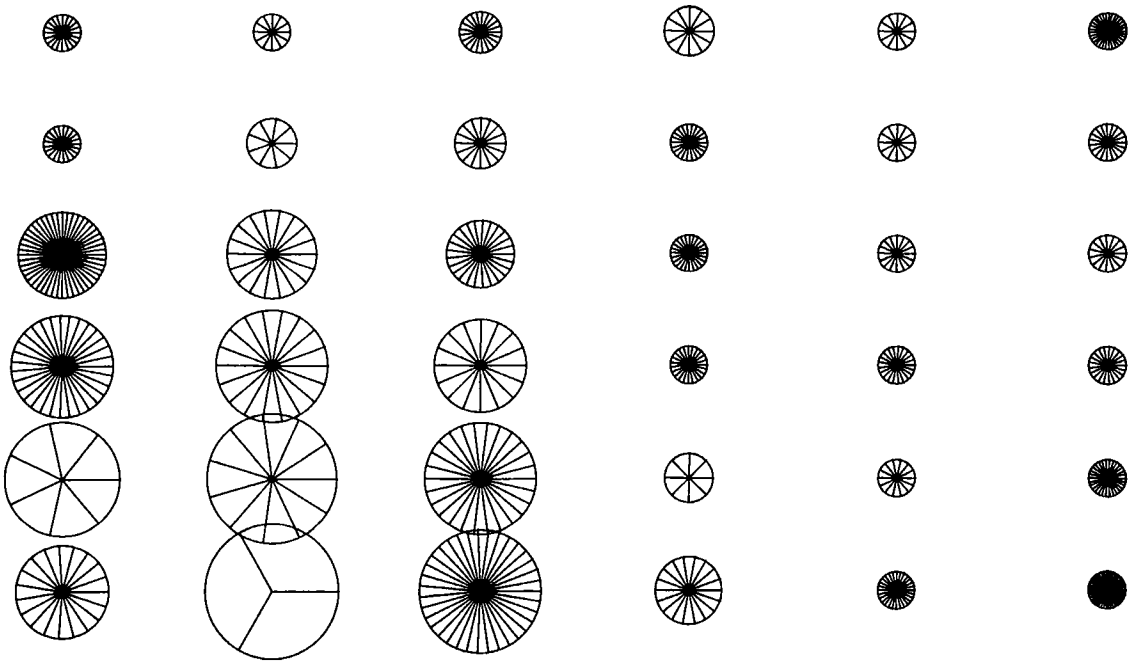
Kohonen's algorithm is especially interesting for visualization, because it simultaneously clusters the data while it arranges the clusters into a planar grid such that neighboring clusters are similar. The resulting grid of clusters is then well-suited for various types of visualization [Ferran and Ferrara 92; Hudson et. al. 89].

Kohonen's Topology Preserving Maps is an iterative neural network algorithm that assigns cluster centroids to a predefined grid of clusters. It works by selecting a training example at random, and then moving the closest cluster centroid *and its neighboring centroids in the grid* a step toward the training example. For this experiment we used 5000 iterations to cluster the 3000+ cases into a 6 by 6 grid (represented by 36 centroids).

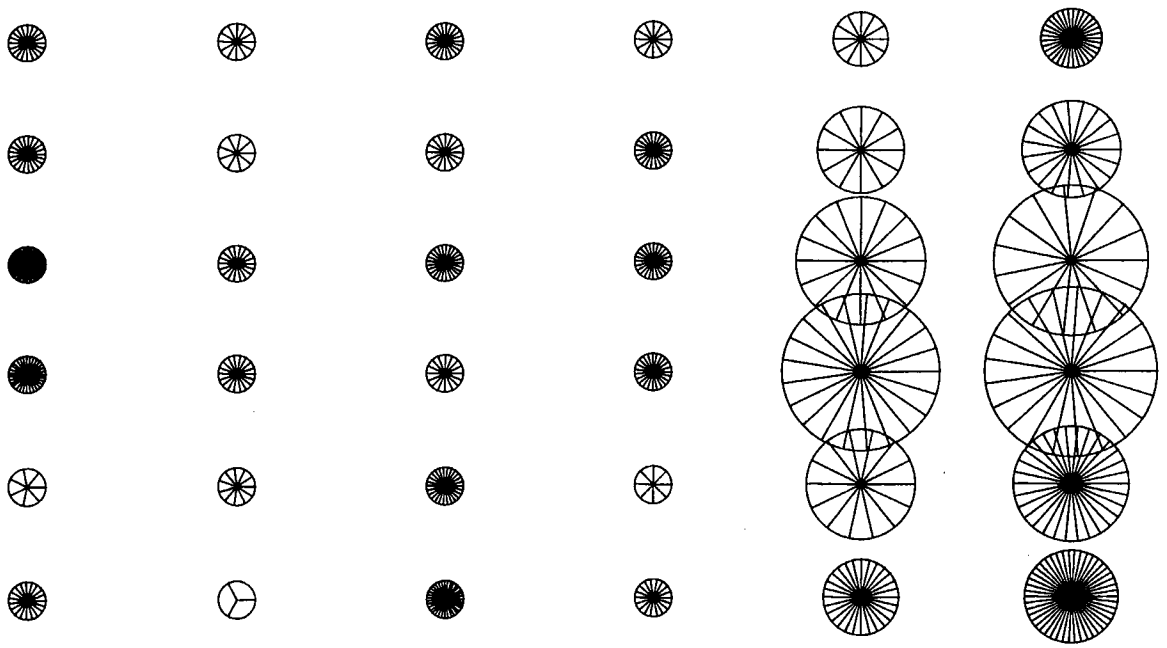
Query Visualization

Once centroids have been determined, we can "query" the clusters using a single feature (e.g. "total insulin dose") or any subset of features. We simply form a query vector by setting features of interest to 1.0 and others to 0, take dot products with all cluster centroids, and finally display this information. More generally, we can select a subset of cases of interest and sum their individual case vectors to form a query vector, or use a cluster centroid as a query vector to identify similar clusters.

For visualization purposes, we can display clusters as circles where each *diameter* corresponds to the dot product of the query vector with that corresponding cluster. Each circle contains "spokes" corresponding to individual patients or groups of patients. (In the example, each spoke represents 5 patients.) In an interactive environment,



Probe: /users/mfj/UL/kohonen/probes/lomeangl.top.probe



Probe: /users/mfj/UL/kohonen/probes/himeangl.top.probe

it is possible to "browse" clusters by clicking on individual cases or small groups, as represented by the spokes.

It is also possible to "label" regions of the grid according to the queries that make them "light up." this produces a set of characterizations for clusters according to those regions in which they appear.

Query Creation

Several strategies can be used to create probes to query the cluster array. Unit vectors on the feature axes can highlight clusters based on single parameters. One or more example vectors may be selected and combined to form more complex queries. We constructed queries by both methods.

Queries based on selected example vectors were formed by sorting the normalized input dataset using one parameter as a sort key, then combining the first or last N vectors as a probe, where N was usually 100 (out of 3721). When these probes were compared to unit vectors with a single nonzero parameter, the example-based queries usually showed more concentrated cluster groups with larger dot products, but both produced qualitatively similar results.

The two figures are plots created with query vectors based on lowest mean blood glucose measurements (top) and highest mean blood glucose measurements (bottom). This illustrates the general properties of the cluster set, using segregation on glucose level as an example. Other related probes were consistent with this topology.

Queries based upon a single patient may also be informative. When the clustering results in clinically relevant subdivisions within a population, this gives an automatic patient classification tool.

Software Environment

All programming was done in Dynamic C++ using BTL (Belmont Toolkit Language). This is a simplified version of C++ with convenient graphics libraries, as well as an integrated run-time debugger, class browser, and automatic garbage collector for quick prototyping. Although Dynamic C++ is an interpreted language, it was easily able to compute clusters for the 3000+ training examples. Code development and graphics interfaces were greatly aided by the built-in tools and libraries.

Prediction

It is also possible to use the expert-representation feature vectors for prediction. For example, we might be interested in predicting pre-dinner blood glucose levels or possibly hypoglycemic episodes. Because the data has been represented by feature vectors, we can use standard neural network algorithms (or other machine learning techniques) for such tasks.

3 Discussion

The combination of expert representations, Kohonen-style displays, and the notion of querying clusters gives an appealing approach to visualization and browsing clinical data.

The roles of these procedures are important. The expert representation makes the data *accessible* to machine learning algorithms, the clustering *arranges data* and *reduces dimensionality*, and the query *focuses* the output on similar cases of interest. In a clinical setting the clustering corresponds to *identifying* classes or variants of a disease or syndrome and to *assigning* patients to a particular cohort based on their symptoms.

Although we decided to concentrate on cluster-based visualization in this paper,

the same data can be used for predicting clinical outcomes by neural networks, and subsequent alarming/reminding based upon those predictions.

References

Ferran, EA and Ferrara, P (1992). "Clustering proteins into families using artificial neural networks." CABIOS 8(1): 39-44.

Gallant, S. *Neural Network Learning and Expert Systems*, MIT Press, 1993.

Hudson, B, Livingstone, DJ, et al. (1989). "Pattern recognition display methods for the analysis of computed molecular properties." Journal of Computer-Aided Molecular Design **3**: 55-65.

Kohonen, T (1982a). "Clustering, taxonomy, and topological maps of patterns." Proceedings of the 6th International Conference on Pattern Recognition October 1982: 114-128.

Kohonen, T (1982b). "Self-organized formation of topologically correct feature maps." Biol Cybern 43: 59-69.

Kohonen, T (1988). Self-Organization and Associative Memory, 2nd Ed. Berlin, Springer-Verlag.

Spilker, B., Crusan, C., Pool, J., Russell, C. and Fram, D. *New Software Technology for Visualizing Clinical Trial Data* . Drug News & Perspectives, Vol. 5, No. 5, June 1992, 298-305.