The Neighborhood Graph for Clinical Case Retrieval and Decision Support within Health-e-Child CaseReasoner

Alexey Tsymbal, Gabor Rendes, Martin Huber
Corporate Technology Division
Siemens AG, Erlangen, Germany
{alexey.tsymbal; gabor.rendes;
martin.huber}@siemens.com

Shaohua Kevin Zhou
Integrated Data Systems Department
Siemens Corporate Research
Princeton, NJ, USA
kzhou@scr.siemens.com

Abstract

In the context of the EU FP6 project Health-e-Child, a Grid-based healthcare platform for European paediatrics is being developed. The basic philosophy behind the design of CaseReasoner, a similarity search based decision support and knowledge discovery system we are developing for Health-e-Child, is to provide a clinician with a flexible and interactive tool to enable operations such as data filtering and similarity search over a Grid of clinical centres, and also to facilitate the exploration of the resulting sets of clinical records regardless of their geographical location. In order to visualize patient similarity, besides the more orthodox heatmaps and treemaps a novel technique based on neighborhood graphs is being developed, which is in the focus of the present paper. For similarity search on distributed biomedical data, besides the canonical distance functions novel techniques for learning discriminative distance functions are also made available to the clinician. The use of distance learning techniques in combination with the patient similarity visualization modules of CaseReasoner contributes to making it a powerful tool for clinical knowledge discovery and decision support in various classification contexts; it helps to combine the power of strong learners with the transparency of case retrieval and nearest neighbor classification.

1 Introduction

There is growing interest in the use of computer-based clinical decision support systems (DSSs) to reduce medical errors and to increase health care quality and efficiency [Berlin *et al.*, 2006]. Clinical DSSs vary greatly in design, functionality, and use. According to the reasoning method used in clinical DSS, one important subclass is that of Case-Based Reasoning (CBR) systems – systems which have reasoning by similarity as the central element of decision support [Berlin *et al.*, 2006; Nilsson and Sollenborn, 2004].

One reason for the slow acceptance of CBR systems in biomedical practice is the especial complexity of clinical data and the resulting difficulty in defining a meaningful distance function on them and adapting the final solution [Schmidt and Vorobieva, 2005]. Another commonly reported reason for the relatively slow progress of the field is the lack of transparency and explanation in clinical CBR. Often, similar patients are retrieved and their diagnoses are presented, without specifying why and to what extent the patients are chosen to be similar and why a certain decision is suggested. We believe that, one way to approach this problem is to better visualize the underlying inter-patient similarity, which is the central concept of any clinical CBR.

In known CBR systems the visualization is usually limited with the visualization of case solutions and not case similarity [Mullins and Smyth, 2001]. To solve the problems described above, we introduce a novel technique for visualizing patient similarity, based on neighborhood graphs, which can be helpful in clinical knowledge discovery and decision making. Besides, we consider two related techniques for learning discriminative distance functions, which when used in combination with the neighborhood graphs can make them a powerful and flexible tool for clinical decision making in different classification contexts.

In this paper we introduce a novel technique for visualizing patient similarity, based on neighborhood graphs; we also discuss the architecture of our implementation within the Health-e-Child DSS CaseReasoner and in particular the related techniques for learning discriminative distance function. The main advantage of the suggested technique is that the decision support becomes *transparent*. The nearest cases and the underlying similarity used for decision making can easily be visualized with the three types of neighborhood graphs. Moreover, after replacing the commonly used "black box" classification with distance function learning and case retrieval, the accuracy of classification usually remains same or even becomes better, and there appears a possibility to visualize the nearest cases of suggested class (say, malignant) and nearest cases of the other class (say, benign), in order for the user (clinician) to analyse and double-check the decision suggested.

The similarity search-based clinical knowledge discovery and decision support system CaseReasoner, besides neighborhood graphs also uses treemaps [Shneiderman, 1992] and heatmaps in order to better represent inter-patient similarity [Tsymbal *et al.*, 2007a]. In particular, the treemap and the heatmap in CaseReasoner represent a hierarchical clustering of patients obtained based on a

certain distance function defined by a clinician e.g. via a set of data attributes of interest or a distance function previously learnt for a certain classification context.

The work in our study has been performed as part of the Health-e-Child (HeC) project. HeC is an EU-funded Framework Programme 6 (FP6) project, which was started in 2006, and aims at improving personalized healthcare in selected areas of paediatrics, particularly focusing on integrating medical data across disciplines, modalities, and vertical levels such as molecular, organ, individual and population. The project of 14 academic, industry, and clinical partners aims at developing an integrated healthcare platform for European paediatrics while focusing on some carefully selected representative diseases in three different categories; paediatric heart diseases, inflammatory diseases and brain tumours. The material presented in this paper contributes to the development of decision support facilities within the platform prototype which provide the clinicians with tools to easily retrieve and navigate patient information and help visualizing interesting patterns and dependencies that may lead, besides personalized decision making concerning appropriate treatment, to establishing new clinical hypotheses and ultimately discovering novel important knowledge.

The paper is organized as follows. In Section 2 the technique of patient similarity visualization based on neighborhood graphs is considered, our implementation of it is discussed and a few examples are given. Section 3 presents techniques for learning discriminative distance functions which can be used to learn a strong distance function in different contexts and which nicely complements the patient similarity visualisation techniques. Section 4 presents the overall architecture of the similarity-search based decision support and knowledge discovery system CaseReasoner, and in Section 5 a few related open issues are discussed with a focus on its evaluation. We conclude in Section 6 with a brief summary, open issues and further research topics.

2 Neighborhood Graphs

2.1 Introduction and Related Work

Neighborhood graphs provide an intuitive way of patient similarity visualization with a node-link entity-relationship representation. There can be distinguished three basic types of neighborhood graphs that can be used to visualize object proximity in DSSs; (1) relative neighborhood graph (RNG), (2) distance threshold graph, and (3) directed nearest neighbor graph. These graphs are studied and applied in different contexts; in particular, as data visualization tools the threshold and nearest neighbor graphs are often used for the analysis of gene expression data in bioinformatics [Zhang and Horvath, 2005; Scharl and Leisch, 2008]. Thus, Zhang and Horvath [2005] study so-called gene co-expression networks, which are represented with the threshold neighborhood graph. Scharl and Leisch [2008] suggest using the nearest neighborhood graph in order to visualize gene clusters.

In a *relative neighborhood graph*, two vertices corresponding to two cases A and B in a data set are connected with an edge, if there is no other case C which is closer to both A and B with respect to a certain distance function d [Toussaint, 1980]:

$$d(A,B) \le \min_{C \ne A,B} \max\{d(A,C), d(B,C)\} \qquad (1)$$

Originally, relative neighborhood graphs were defined for 2D data (planar sets) with the Euclidean distance metric, but later they were generalized and applied to multiple dimensions and other distance functions [Toussaint, 1980; Jaromczyk and Toussaint, 1992; Muhlenbach and Rakotomalala, 2002].

Besides the relative neighborhood graphs we focus on, there are known some other related node-link (graph-based) visualizations of instance proximity. These include the Minimum spanning tree (MST), the Gabriel graph, and the Delanay tessellation [Jaromczyk and Toussaint, 1992]. We believe that out of this family, the relative neighborhood graph is the best candidate to visualize patient proximity in a DSS. The MST has usually too few edges to spot groupings/patterns in the data, while the Gabriel graph and the Delanay tessellation are, vice versa, usually too overcrowded, which becomes a problem with already more than a hundred cases (patients).

A *threshold* graph is simply defined as a graph where two vertices are connected with an edge if the distance between the two corresponding cases is less than a certain threshold. In a *nearest neighbor graph*, each case is connected with one or a set of its nearest neighbors. This graph is usually directed as the relation of being a nearest neighbor is not necessarily symmetric. An important benefit of RNG comparing to the other two graphs is the fact that it is always connected with nodes having a reasonable small degree; it is often planar or close to planar.

In machine learning, neighborhood graphs find various applications, including data clustering, outlier removal, and even supervised discretization [Muhlenbach and Rakotomalala, 2002]. The k-nearest neighbor (k-nn) graph is often used as the base in various approximate nearest neighbor search techniques in high-dimensional spaces, in order to cope with the curse of dimensionality, see [Sebastian and Kimia, 2002; Paredes and Chavez, 2005] for two examples. Besides, neighborhood graphs may serve as a source of measures of complexity for such searching, in order to estimate the costs [Clarkson, 2006]. Another important related branch of research studies how to optimize the process of construction of the neighborhood graph. Thus, Paredes *et al*. [2006] optimize construction of the k-nearest neighbor graph in metric spaces and achieve empirically around $O(n^{1.27})$ complexity in low and medium dimensional spaces and $O(n^{1.9})$ in high dimensional ones. [Jaromczyk and Toussaint, 1992] review algorithms for reducing the complexity of constructing an RNG, which is $O(n^3)$ in general, in low-dimensional spaces.

Besides machine learning and similarity search optimization, in other domain areas other, more exotic applications may also be found. In [Marcotegui and Beucher, 2005], for example, the minimum spanning tree of a neighborhood graph was applied to contrast-based hierarchical image segmentation. In [Li and Hou, 2004] a directed relative neighborhood graph and directed minimum spanning tree are successfully applied to topology control, in order to create a power-efficient network topology in wireless multi-hop networks with limited mobility.

Fig. 1 below presents an example of a neighborhood graph constructed for the Leukemia public gene expression data set (available at www.upo.es/eps/aguilar/ datasets.html) within CaseReasoner. RNG for a set of 72 samples representing healthy (blue) and diseased (red) pa-

tients is shown. The underlying distance function is the intrinsic Random Forest distance. Leave-one-out accuracy for this problem is as high as 98%. Such a graph provides a powerful tool for knowledge discovery and decision making in the considered domain (e.g., by placing and displaying the gene expression sample of a new patient with unknown diagnosis in such a graph).

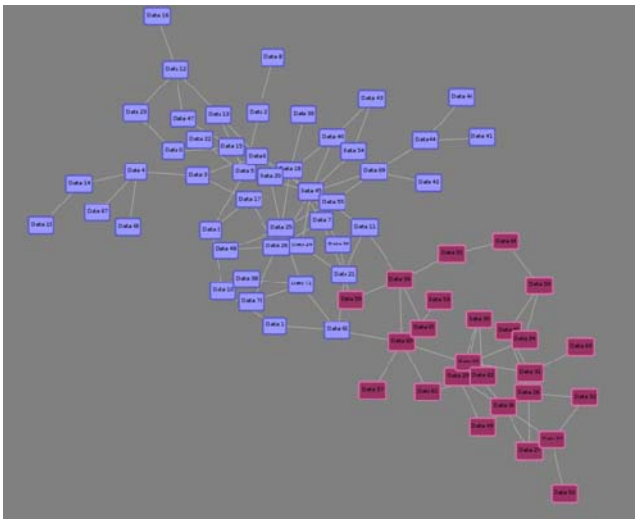

Figure 1 A relative neighborhood graph for the Leukemia dataset

2.2 Functionality and GUI

In our toolbox for visualization, navigation and management of the three neighborhood graphs introduced above, which is being developed as a part of the clinical DSS CaseReasoner, we implemented the following functionality:

- node coloring, to represent numeric and nominal attributes;
- node filtering, according to attribute values in the patient record;
- edge coloring and filtering, according to the underlying distance;
- graph (hierarchical) clustering into an arbitrary number of components including a panel for clustering tree navigation on the graph;
- reconfigurable tooltips displaying clinical data from the patient record and images;
- nearest neighbor classification and regression performance visualization for each node, for a selected class attribute and a certain similarity context;
- image visualization within the nodes of the graph (e.g. meshes corresponding to the pulmonary trunk of the patient can be displayed).

Besides clinical data and patient similarities, the neighborhood graphs are nicely suitable for displaying images corresponding to patients. The same operations can still be used as for usual graphs (graph clustering, node coloring and filtering, edge coloring and filtering, etc.); also the images (e.g., meshes) can be scaled and rotated. In Fig. 2 below the meshes corresponding to the pulmonary trunks of the patients are displayed within the nodes of a graph displaying a cohort of HeC cardiac patients, and a sketch of GUI of the neighorhood graph module is shown, including the toolbar with access to the basic operations on the graph such as coloring, filtering and clustering, a pop-up graph settings control panel and a status bar with basic information about the currently displayed graph. The interactive graph navigation is implemented using the *Prefuse* open source data visualization toolkit as the core [Heer *et al.*, 2005].

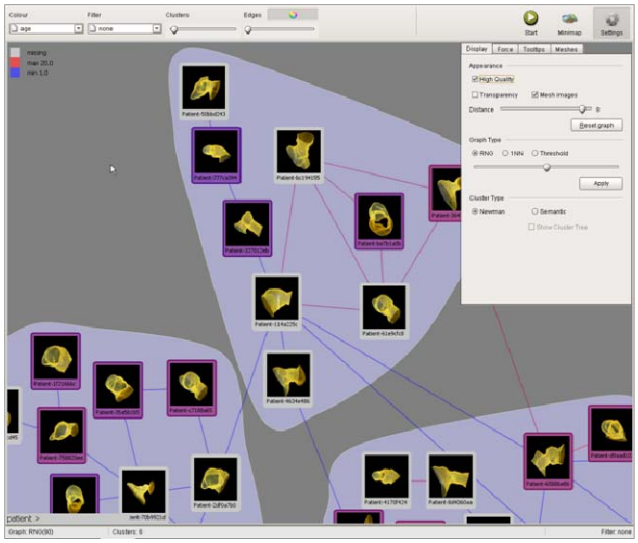

Figure 2 GUI and image visualization within the neighbor graph

The GUI for the basic functionality was designed to be intuitive and straightforward. E.g., if node coloring option is activated, a small legend tells which colors represent the maximum and minimum value or in the case of a nominal value, the range of displayed feature values. In the case of node filtering, the filtered values or ranges are also displayed. Edge coloring represents the distances between the patients with a color range, from blue (weak connection) to red (strong connection). The edge filtering functionality removes a given percent of the weakest connections from the graph, so only the more relevant connections will be remaining.

Besides coloring each node according to a selected attribute, the node color may also be selected to represent the predictive performance of the current similarity context with respect to a certain nominal or numeric feature. In particular, for every node, the leave-one-out estimate of margin or 0/1 loss function with nearest neighbor classification can be displayed, or the leave-one-out estimate of the absolute error with nearest neighbor regression for numeric attributes can be visualized.

As the underlying distance function used for the construction of the neighborhood graph, two options may be considered. The first is to use a certain canonical distance function, such as the well known Euclidean metric, with a possibility to change the influence/importance of every contributing feature. However, due to the inherent complexity of medical data, the canonical distance function may not always represent the true similarity among patients, and thus be suboptimal for decision support. A solution to this issue is to learn a strong distance function for a given classification context under consideration using the available labeled data for training.

Two techniques for learning discriminative distance functions were implemented by us and evaluated; learning from equivalence constraints in the product or difference

spaces and the intrinsic Random Forest (RF) distance. Our experiments confirm that both techniques demonstrate competitive performance with respect to the plain learning, and are suitable to be used in combination with the neighborhood graph visualization. The intrinsic RF distance is proven to be more robust overall in our experiments, although finding suitable parameters for learning from equivalence constraints may still be competitive. A thorough introduction to the techniques for learning discriminative distance functions is given in Section 3. We use both the canonical and the discriminative distance functions for constructing the graphs, and in the case of the latter one, customized models can be generated, which can be stored and retrieved later on, in order to specify the similarity context of interest.

For clustering the neighborhood graphs, we use the following two algorithms; (1) the Girvan and Newman's algorithm for graph clustering which is often used for clustering of complex social and biological networks [Girvan and Newman, 2002], (2) top-down induction of a semantic clustering tree (in the original feature space), the goal of which is to provide every cluster with a semantic description that can be inspected by a clinician and may carry important information.

The *semantic clustering* algorithm was developed specifically for CaseReasoner; we could not find a same algorithm already described in the literature, although it is simple and many similar approaches do exist. The related algorithms differ in the structure of the generated cluster descriptions. Our main intention was to provide a *tree* with semantic splits in the nodes that could be used in order to navigate the hierarchical clustering generated and explore the clusters. In order to generate the tree, we use a similar top-down decision tree induction procedure which is often used for supervised learning (e.g. the C4.5 decision tree). Similar to the supervised case, all possible univariate semantic splits are examined in each node (such as '*gender=F'* or '*age<2'*). As the criterion to find the best split, the ratio of between-cluster variance to the within-cluster variance is used. Variance is defined in terms of the current similarity context. If it is specified as a set of features of interest, then the variance can be calculated directly on them. If a customized distance function is loaded, then the variance is represented via distances between a pair of within- and between- cluster cases. According to the first feedback of clinicians regarding the implemented semantic clustering algorithm, the generated tree often contains useful information and may serve as a certain description of the current similarity context.

In Fig. 3 GUI of the neighborhood graph module within the HeC DSS CaseReasoner is shown. The graph shown displays the semantic clustering of a cohort of cardiac patients according to the currently selected similarity context, and node color represents blood pressure for corresponding patient. The pop-up control panel in the upper right corner is used for navigation over the current clustering tree; each cluster can be centered and highlighted and split further by clicking on the corresponding node in the tree in this panel. The panel on the left to the graph navigation panel is the patient record navigation panel which is used in order to browse and compare feature values and their place in the general feature distribution for the current and most similar patient.

3 Learning Discriminative Distance Functions

There are several reasons that motivate the studies in the area of learning distance functions and their use in practice [Bar-Hillel, 2006]. First, learning a distance function helps to combine the power of strong learners with the transparency of nearest neighbor classification. Moreover, learning a proper distance function was shown to be especially helpful for high-dimensional data with many correlated, weakly relevant and irrelevant features, where most traditional techniques would fail. Also, it is easy to show that choosing an optimal distance function makes classifier learning redundant. Next, learning distance functions breaks the learning process into two sequential steps (distance learning followed by classification or clustering), where each step requires search in a less complex functional space than in the immediate learning. Moreover, it fosters the creation of more modular and thus more flexible systems, supporting component reuse. Another important benefit is the opportunity for inductive transfer between similar tasks; this approach is often used in computer vision applications; see e.g. [Mahamud and Hebert, 2003].

Historically, the most popular approach in distance function learning is Mahalanobis metric learning, which has received considerable research attention but is however often inferior to many non-linear and non-metric distance learning techniques. While distance metrics and kernels are widely used by various powerful algorithms, they work well only in cases where their axioms hold [Hertz, 2006]. For example, in [Jacobs *et al*., 2000] it was shown that distance functions that are robust to outliers and irrelevant features are non-metric, as they tend to violate the triangular inequality. Human similarity judgements were shown to violate both the symmetry and triangular inequality metric properties. Moreover, a large number of hand-crafted context-specific distance functions suggested in various application domains are far from being metric. Our focus is thus on techniques for learning non-linear and non-metric discriminative distance functions, two important representatives of which are considered in sub-sections below.

More than in any other research domain, the problem of learning a better distance function lies in the core of research in *computer vision* [Bar-Hillel, 2006]. Different imaging applications have been considered, including image retrieval (with facial images, animal images, hand images, and American Sign Language images), object detection (indoor object detection), motion estimation and image registration; see [Hertz, 2006; Bar-Hillel, 2006] for an in-depth review.

Besides vision, some other domains were also considered including computational immunology, analysis of neuronal data, protein fingerprints, and text retrieval [Hertz, 2006]. Surprisingly, there is relatively few related work in text/document retrieval. One example is [Schulz and Joachims, 2003] which studies the retrieval of text documents from the Web by learning a distance metric from comparative constraints.

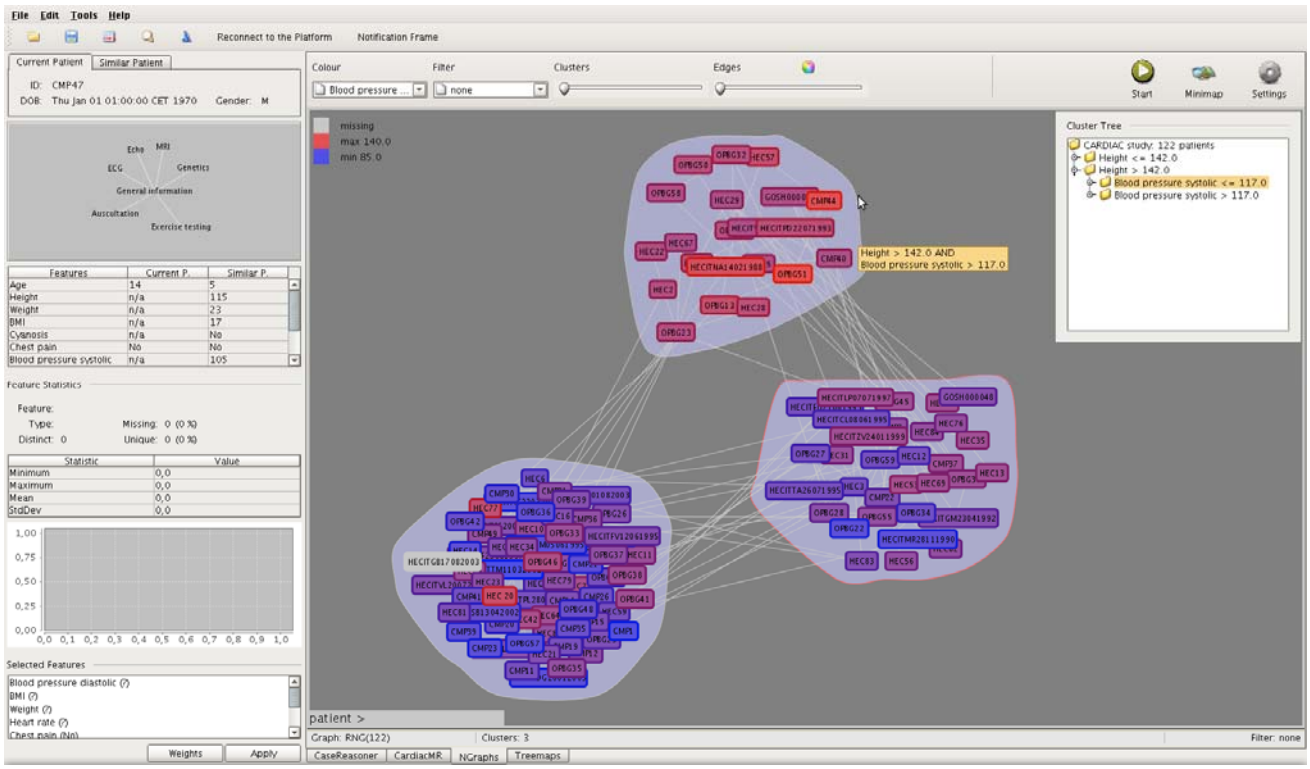

Figure 3 GUI of the neighborhood graph module within HeC CaseReasoner

Figure 4 The HeC CaseReasoner application, the workflow

3.1 Learning from Equivalence Constraints

Usually, equivalence constraints are represented using triplets (x_1, x_2, y), where x_1, x_2 are data points in the original space and $y \in \{+1,-1\}$ is a label indicating whether the two points are similar (from the same class) or dissimilar. Learning from these triples is also often called learning in the *product space* (i.e. with pairs of points as input); see [Hertz *et al*., 2004; Zhou *et al*., 2006] for examples. While learning in the product space is perhaps a more popular form of learning from equivalence constraints, yet another common alternative is to learn in the *difference space*, the space of vector differences; see [Amores *et al*., 2006; Yu *et al*., 2006] for examples. The difference space is normally used with homogeneous high-dimensional data, such as pixel intensities or their PCA coefficients in imaging. While both representations demonstrate promising empirical results in different contexts, there is no understanding which representation is better. No comparison was done so far; usually a single representation for the problem is chosen.

There are two essential reasons that motivate the use of equivalence constraints in learning distance functions; their availability in some learning contexts and the fact that they are a natural input for optimal distance function learning [Bar-Hillel, 2006]. It can be shown that the optimal distance function for classification is of the form $p(y_i \neq y_j \mid x_i, x_j)$. Under the independence and identical distribution (*i.i.d.)* assumption the optimal distance measure can be expressed in terms of generative models $p(x \mid y)$ for each class as follows [Mahamud and Hebert, 2003]:

$$p(y_i \neq y_j \mid x_i, x_j) = \sum p(y \mid x_i)(1 - p(y \mid x_j)) \quad (2)$$

3.2 The intrinsic Random Forest distance function

For a Random Forest (RF) learnt for a certain classification problem, the proportion of the trees where two instances appear together in the same leaves can be used as a measure of similarity between them [Breiman, 2001]. For a given forest f the similarity between two instances x_1 and x_2 is calculated as follows. The instances are propagated down all K trees within f and their terminal positions $\boldsymbol{z}$ in each of the trees ($\boldsymbol{z_1}=(z_{11},...,z_{1K})$ for x_1, similarly $\boldsymbol{z_2}$ for x_2) are recorded. The similarity between the two instances then equals to (I is the indicator function):

$$S(x_1, x_2) = \frac{1}{K} \sum I(z_{1i} = z_{2i}) \quad (3)$$

Similarity (2) can be used for different tasks related to the classification problem. Thus, Shi and Horvath [2006] successfully use it for hierarchical clustering of tissue microarray data. First, unlabeled data are expanded with a synthetic class of evenly distributed instances, then a RF is learnt and the intrinsic RF similarities are determined as described above and clustered. The resulting clusters are shown to be clinically more meaningful than the Euclidean distance based clustering with regard to post-operative patient survival.

Interesting is that using this similarity for the most immediate task, nearest neighbor classification, is rather uncommon, comparing to its use for clustering. In one of related works, [Qi *et al*., 2005], it is used for protein-protein interaction prediction, and the results compare favourably with all previously suggested methods for this task.

The intrinsic RF distance is rather a "dark horse" with respect to learning from equivalence constraints. The number of known applications for it is still limited; perhaps, the most successful application is clustering genetic data, [Shi and Horvath, 2006]. Works on learning equivalence constraints never consider it as a possible alternative. In general, we believe that the circle of applications both for distance learning from equivalence constraints (which is currently applied nearly solely to imaging problems) and for the intrinsic RF distance is still, undeservedly, too narrow and may and should be expanded.

4 CaseReasoner: A Framework for Medical CBR

The basic philosophy behind the design of the CaseReasoner is to provide clinicians with a flexible and interactive tool to enable operations such as data filtering and similarity search over a Grid of clinical centres (following the formerly introduced information retrieval paradigm), and also to facilitate the exploration of the resulting data sets. The aim is to let clinicians explore and compare the patients' records regardless of his/their geographical location, and to visualize their place in the distribution of both the whole population of patients, as well as in the distribution of its semantic subsets.

The selected visualization techniques are implemented to display and navigate through the results of similarity searches in the CaseReasoner. The distance function for similarity search is defined based on a similarity context, which is a subset of features of interest defined by the clinician. The features for each problem domain are organized into a so-called feature ontology, which represents the relationships between features [Tsymbal *et al*., 2007b]. Similar cases are found both in the whole Integrated Case Database (ICD) over the Grid, and in some subsets of interest (e.g. high-grade tumours, males, a certain node in the Grid, etc), defined in the form of a simple filter. For each patient in the ICD, it is possible to visualize and compare related images from the patient history, thanks to the Gateway's abstracted accesses to backends, storage elements and file catalogs. In combination with the basic feature statistics, class distribution histograms and the scatter plots for the ICD under study, this will be a universal tool for Grid-based decision making in the diseases covered by HeC. Indeed, having a number of clinical centres connected and sharing their data gives the CaseReasoner significant added value. Not only can the CaseReasoner benefit from larger samples but also part of its reasoning logic can be made reusable and delegated to the Grid, with the complexity that it implies.

In short and as illustrated in Figure 4, after selecting a search context (1), the clinician can view basic statistics of the retrieved cases (2) as well as visualize them utilizing neighborhood graphs (3a), treemaps (3b) and heatmaps (3c).

In the development of the general code structure for the CaseReasoner framework we followed a less strict variation of the Presentation-Abstraction-Control pattern (PAC) [Coutaz, 1987]. The main idea behind that pattern is a hierarchical structure of 'agents', where the core of the framework acts as a top level agent, and the modules

are subordinate agents. The core and the modules communicate with each other only through their Controller part. This way the flow of data and control remains clear: the general, essential workflow is managed by the core Controller, while the modules can handle the inner business logic of their own. They can request general data manipulating operations through the core, and they are notified about every relevant change as well.

In such a flexible framework any module can be added and removed easily, their development process is fully independent from the whole framework, whose API is well defined and easy to use. Currently CaseReasoner provides five data visualisation modules: the NGraphs module for neighborhood graphs presented above, the TreeMaps, the Heatmapper module, the Patient Panel, to display and compare values and statistics of single patient records, and the CardiacMR module which can visualise and navigate cardiac imaging data related to a selected patient.

5 Discussion

In the evaluation of the HeC platform and in particular the DSS CaseReasoner we follow the "Multi-dimensional In-depth Long-term Case studies" (MILCs) paradigm proposed in [Shneiderman and Plaisant, 2006]. In the MILCs concept the *multi-dimensional* aspect refers to using multiple evaluation techniques including observations, interviews, surveys, as well as automated logging to assess user performance and interface efficacy and utility [Shneiderman and Plaisant, 2006]. In the context of our project, mostly observations, interviews and questionnaires were used so far in order to obtain feedback and assess user satisfaction with the platform. The *in-depth* aspect is the intense engagement of the researchers (the IT experts within Health-e-Child) with the expert users (clinical partners within the project) to the point of becoming a partner or assistant. *Longterm* refers to longitudinal studies that begin with training in use of a specific tool through proficient usage that leads to strategy changes for the expert users. The initial phase of our evaluation has started already in mid 2006, with the start of the project, by demonstrating the preliminary versions of the prototypes for certain platform modules (not yet fully functional) to the clinicians and collecting their requirements to the extension and revision of the tools in so-called knowledge elicitation sessions. The main task of this phase was for the IT partners to better understand the problem domain and the needs of clinicians and develop the tools to fully satisfy these needs. Data collection for the platform has also started in parallel. This iterative evaluation and development phase, when the first fully functional platform prototype was ready, has been gradually replaced (by mid 2008) with the training phase. The main task of this phase which consists of a series of on-site training sessions and will last till the end of the project (April 2010) is to train all the participating clinicians (from the four hospitals in the UK, Italy and France) to use the platform with the collected patient records on the premises of the hospitals and to obtain their extensive feedback in order to better evaluate the platform and fix its discovered deficiencies if needed. Our ultimate goal, which is still for us to achieve is to improve healthcare quality and efficiency and reduce costs for the participating hospitals. *Case studies* refers to the detailed reporting about a small number of individuals working on their own problems, in their normal environment [Shneiderman and Plaisant, 2006].

Perhaps the main competitor to neighborhood graphs as a tool for visualizing patient similarity is heatmaps, which are well known and often used by clinical researchers, in particular by geneticists. In comparison to heatmaps, as follows also from the feedback obtained from partner clinicians in our project, neighborhood graphs possess a number of advantages. In particular, they are easier to read with the more intuitive node-link entity-relationship representation, they allow visualizing additional features or even image thumbnails at nodes, and they have a flexible layout allowing to naturally visualize clusters, enlarge nodes, and filter our a set of nodes and edges.

6 Conclusions

In this paper we introduced a novel technique for visualizing patient similarity, based on neighborhood graphs, which could be helpful in clinical decision making; we also discussed the architecture of our implementation within the Health-e-Child DSS CaseReasoner and in particular the related techniques for learning discriminative distance function. The main advantage of the suggested technique is that the decision support becomes *transparent*; the power of strong machine learning techniques via discriminative distance learning is combined with the transparency of nearest neighbor classification.

An important issue of our ongoing work is a better acquaintance of partner clinicians with the considered neighborhood graph module in the framework of CaseReasoner, and its evaluation in the context of different data classification and decision support tasks. An important issue of our ongoing work with distance learning is the study of the use of online learning techniques for learning from equivalence constraints and in particular the incrementalization of Random Forests, in order to speed up learning in the product space.

Acknowledgements

This work has been partially funded by the EU project Health-e-Child (IST 2004-027749). The authors wish to acknowledge support provided by all the members of the Health-e-Child consortium in the preparation of this paper.

References

[Amores *et al.*, 2006] Jaume Amores, Nicu Sebe, Petia Radeva. Boosting the distance estimation. Application to the *k*-nearest neighbor classifier. *Pattern Recognition Letters* 27, 2006, 201-209.

[Bar-Hillel, 2006] Aharon Bar-Hillel. *Learning from Weak Representations Using Distance Functions and Generative Models*. Ph.D. Thesis, Dept. Comp. Sci., Hebrew Univ. of Jerusalem, 2006.

[Berlin *et al.*, 2006] Amy Berlin, Marco Sorani, Ida Sim. A taxonomic description of computer-based clinical decision support systems. *J. of Biomedical Informatics*, 39(6), 2006, 656-667.

[Breiman, 2001] Leo Breiman. Random Forests. *Machine Learning* 45 (1), 2001, 5-32.

[Clarkson, 2006] Kenneth L. Clarkson. Nearest-neighbor searching and metric space dimensions. (Survey). In: G. Shakhnarovich, T. Darell, P. Indyk (ed.), Nearest-

Neighbor Methods foe Learning and Vision: Theory and Practice, MIT Press, 2006, 15-59.

[Coutaz, 1987] Joëlle Coutaz. PAC: an implementation model for dialog design. In: H-J. Bullinger, B. Shackel (ed.), *Proc. Interact'87 Conference,* North-Holland, 1987, 431-436.

[Heer *et al.*, 2005] Jeffrey Heer, Stuart K. Card, James A. Landay. Prefuse: a toolkit for interactive information visualization. In: *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems, CHI'05*, ACM Press, 2005, 421-430.

[Hertz *et al.*, 2004] Tomer Hertz, Aharon Bar-Hillel, Daphna Weinshall. Learning distance functions for image retrieval. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition,* 2004.

[Hertz, 2006] Hertz Tomer. *Learning Distance Functions: Algorithms and Applications*. Ph.D. Thesis, Dept. Comp. Sci., Hebrew University of Jerusalem, 2006.

[Jacobs *et al.*, 2000] David W. Jacobs, Daphna Weinshall, Yoram Gdalyahu. Classification with non-metric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(6), 2000, 583-600.

[Jaromczyk and Toussaint, 1992] Jerzy W. Jaromczyk and Godfried T. Toussaint. Relative neighbourhood graphs and their relatives. In: *Proc. IEEE*, 80(9), 1992, 1502-1517.

[Li and Hou, 2004] Ning Li and Jennifer C. Hou. Topology control in heterogeneous wireless networks: problems and solutions. In: *Proc. 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Infocom'04*, IEEE, 2004.

[Mahamud and Hebert, 2003] Shyjan Mahamud, Martial Hebert. The optimal distance measure for object detection. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition,* 2003.

[Marcotegui and Beucher, 2005] Beatriz Marcotegui and Serge Beucher. Fast implementation of waterfall based on graphs. In: *Proc. 7th Int. Symposium on Mathematical Morphology, Computational Imaging and Vision*, Vol. 30, Springer, 2005, 177-186.

[Muhlenbach and Rakotomalala, 2002] Fabrice Muhlenbach and Ricco Rakotomalala. Multivariate supervised discretization, a neighbourhood graph approach. In: *Proc. IEEE International Conference on Data Mining, ICDM'02*, IEEE Computer Society, 2002, 314-321.

[Nilsson and Sollenborn, 2004] Markus Nilsson, Mikael Sollenborn. Advancements and trends in medical case-based reasoning: an overview of systems and system development. In: *Proc. 17th Int. FLAIRS Conf. on AI, Special Track on CBR*, AAAI Press, 2004, 178-183.

[Paredes and Chavez, 2005] Rodrigo Paredes, Edgar Chavez. Using the k-nearest neighbor graph for proximity searching in metric spaces. In: Proc. SPIRE'05, LNCS 3772, 2005, 127-138.

[Paredes *et al.*, 2006] Rodrigo Paredes, Edgar Chavez, Karina Figuero, Gonzalo Navarro. Practical construction of k-nearest neighbor graphs in metric spaces. In: 5th Int. Workshop on Experimental Algorithms WEA'06, LNCS 4007, Springer, 2006, 85-97.

[Qi *et al.*, 2005] Yanjun Qi, Judith Klein-Seetharaman, Ziv Bar-Joseph. Random Forest similarity for protein-protein interaction prediction from multiple sources. In: *Proc. Pacific Symp. on Biocomputing*, 2005.

[Scharl and Leisch, 2008] Theresa Scharl and Friedrich Leisch. Visualizing gene clusters using neighborhood graphs in R. Tech. Report 16, Dept. of Statistics, Uni. of Munich, 2008. Available at http://epub.ub.uni-muenchen.de/2110/1/tr016.pdf.

[Schulz and Joachims, 2003] Matthew Schulz, Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in Neural Information Processing Systems*, NIPS 16, 2003.

[Schmidt and Vorobieva, 2005] Rainer Schmidt and Olga Vorobieva. Adaptation and medical case-based reasoning focusing on endocrine therapy support. In: *Proc. Int. Conf. on AI in Medicine, AIME'05*, LNCS, Vol. 3581, Springer, 2005, 300-309.

[Sebastian and Kimia, 2002] Thomas B. Sebastian, Benjamin B. Kimia. Metric-based shape retrieval in large databases. In: Proc. 16th Int. Conf. on Pattern Recognition, Vol. 3, 291-296.

[Shi and Horvath, 2006] Tao Shi, Steve Horvath. Unsupervised learning with Random Forest predictors. *Computational and Graphical Statistics* 15 (1), 2006, 118-138.

[Shneiderman, 1992] Shneiderman B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1), 1992, 92-99.

[Shneiderman and Plaisaint, 2006] Ben Shneiderman, Catherine Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In: *Proc. AVI Workshop on Beyond time and errors: novel evaluation methods for information visualization, BELIV*, ACM Press, 2006, 1-7.

[Toussaint, 1980] Godfried T. Toussaint. The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, 12(4), 1980, 261-268.

[Tsymbal *et al.*, 2007a] Alexey Tsymbal, Martin Huber, Sonja Zillner, Tamas Hauer, Shaohua K. Zhou. Visualizing patient similarity in clinical decision support. In: A. Hinneburg (ed.), *LWA 2007: Lernen - Wissen - Adaption, Workshop Proc.*, Martin-Luther-University Halle-Wittenberg, 2007, 304-311.

[Tsymbal *et al.*, 2007b] Alexey Tsymbal, Sonja Zillner, Martin Huber. Ontology- supported machine learning and decision support in biomedicine. In: *Proc. 4th Workshop on Data Integration in the Life Sciences, DILS'07*, LNBI, Springer, 2007, 156-171.

[Yu *et al.*, 2006] Jie Yu, Jaume Amores, Nicu Sebe, Qi Tian. Toward robust distance metric analysis for similarity estimation. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition, CVPR, 2006.*

[Zhang and Horvath 2005] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis, *Statistical Applications in Genetics and Molecular Biology, 1 (17), 2005.*

[Zhou *et al.*, 2006] Shaohua K. Zhou, Jie Shao, Bogdan Georgescu, Dorin Comaniciu. Boostmotion: boosting a discriminative similarity function for motion estimation. In: *Proc. Int. Conf. on Computer Vision and Pattern Recognition,* 2006.