A CBR Dominant multi strategy design:

Getting the most from Intelligent Systems

A. Sunny Ochi-Okorie

Inference Corporation,

100 Rowland Way, Suite 300, Novato, CA 94945

sunny.okorie@inference.com

Abstract

A Case Based Reasoning (CBR) *dominant* diagnostic system is presented which collaborates with other intelligent problem solving methodologies. Two of the contributing elements, namely the *Matched Vector Functions (MVF*) and the *Evidence Ratio Factors* (ERF) propose an initial decision following their individual data pre-processing. Further, using the *Singular Value Decomposition* (SVD) method, the proposed solutions are improved and cast as a vote among three main collaborating peers that includes the CBR unit. The paper gives a brief overview of *CBR integration* in a *multi strategy* design, and goes further to suggest how to get the most out of intelligent systems that have been so deployed. The main motivation and goal here is performance improvements in the diagnostic accuracy of the system when compared with actual clinician's diagnostic results, as well as ease of use, and the differential diagnosis feature. Details are discussed in the text and appendix with some results of the capabilities that have benefited our research effort.

1.0 Introduction

CBR integrations with other AI methods promise higher success rates for various degrees of problem solving, reference solutions and related applications. Humans tend to base most decision-making on what can be recalled from past experience. This is fairly common knowledge from cognitive psychology. Often, facts are deliberately neglected in our daily problem solving when we use our sixth sense, our hunch, or professional judgement. Why do we behave in this manner? Why does the sixth sense, memory, or whatever else we may call it play a *dominant* role in our decision making?

CBR has been integrated successfully in many commercial settings for dealing with *Customer Support* services of various kinds. These sites or settings usually have Problem or Change Tracking Systems deployed in the *help-desk* environment or the *World Wide Web* (WWW) *Internet self-service.* In many others, the problem at hand may be to readily find answers to basic questions of a client, or customer, thereby making *Information Retrieval* (IR) a routine activity in the workplace.

A number of robust commercial search engines have been built for especially Internet / WWW searches enabling users to intelligently collect, categorize, and sort data in different ways in order to quickly find what is desired from the massive information on the *Internet.* Some of the search engines have been referred to *search agents* when equipped with features to personalize your search or similar capability is present.

In this paper, we describe new design integration to the application, **TROPIX** (Ochi-Okorie, 1997), which was developed for the diagnosis and treatment of tropical diseases. From this research extension on **TROPIX**, current test results show that combining reasoning paradigms for diagnosis and therapy selection improved our results considerably over what it was from just one or two approaches used in an earlier prototype (from a lower measure of 86% to approximately 98.5% accuracy).

2.0 Review of Related Work

One very important task-driven application that has integrated CBR technology with other computing methods can be found in some Web search engines. Inference Corporation's **INFIND**[1], and its' subsidiary's **ZURFRIDER** have certainly demonstrated competence in very large or *mega* search of data repositories. Using the Internet as the case base, their goal is to rapidly help a user to find the needed information or resource in the shortest possible time. Thus, these search engines collect a user's search description, formulate parallel search queries for different Web search engines, and will target them toward likely Web repositories. Returned results are then organized into neatly sorted groups or categories before presenting them to the user with Web sites showing the best hits coming at the top of the returned list of folders / web site files. **ZURFRIDER** is capable

[1] INFIND is a Web Search tool developed by Inference Corporation, Novato, CA. http://www.inference.com/

of constructing a small set of questions / optional feature selections for the user to further filter the search retrieval on a subsequent search. Many search algorithms use simple character text, or word matching to find required information from a file, folder, database, or some extremely large repository such as the Internet. Simple character or word search is not adequate for an Internet application. However, **INFIND** uses proprietary methods that incorporate matched triagrams (three adjacent characters including white spaces) in its description-based text search.

The integration of *Artificial Neural Network* (ANN) with CBR is currently being explored by a small number of researchers. These include the work of (Lees and Corchado 1997), (Krovvidy and Wee 1993), as well as (Lees, Rees, and Aiken, 1992). These and other research effort combine CBR and other AI methods such as rule-based systems, rough sets, fuzzy sets, Bayesian nets, or multivariate statistical learning methods. Lees et al in their paper (1997) deal with exploratory use of CBR and ANN to improve predictions from both historical data and new data acquired in real time for a sea going vessel. The goal is to continuously generate valid physical parameters in the immediate vicinity of the vessel such as temperature changes, gradients, and other oceanographic information of interest in three dimensions. *CBR dominates* the operation of this intelligent system in that it systematically can retire the ANN component until it is retrained periodically.

In our research, a major goal that cannot be compromised is quality. In the medical field and diagnosis, second opinion is also important to all experts alike since a number of decisions have to be made from often very *fuzzy (noisy) information.* The idea of *differential diagnosis* is well provided for in the **TROPIX** prototype that has already been built and demonstrated (Ochi-Okorie, 1997, 1998). It was related to match ties in diagnoses between disease classes. However, the drive to improve the system by incorporating better methods of reasoning, restructuring the various contributing units, etc. to realize better results continues.

3.0 On-going Work in TROPIX

In **TROPIX**, (Ochi-Okorie 1996, 1997), patient diagnosis with any of the 22 tropical diseases in the system is done initially using heuristic matching algorithms, *MVF* (*Matched Vector Functions*) from multivariate statistics and pattern recognition. Its core algorithm is based on Similarity *Functions* (Joly S. and Le Calve' G, 1994) which were extended to incorporate *dissimilarity analysis.* In the present work, the *Extended Similarity Functions* (*ESF*) of the MVF procedure are used alone to get a diagnosis by computing *ESF* (*disease*, j) = {Σ(*similar features by inner product moment*) + Σ(*absent features in both the new case and Domain Knowledge model, DM*) - Σ(*dissimilar features*)} * k, some domain features-type constant.

In previous work, we used *ERF* (*Evidence Ration Factors*) to confirm a diagnosis or otherwise. The *ERFs* are based on some *priori probabilities* determined from a number of domain experts, but are now subjected to revisions as the system accumulates clinical cases in the present work.

Preliminary diagnoses with the *MVF* algorithm using the *Knowledge Base* and the new patient's clinical data were yielding between 86% to 90% success rates. Our goal was to improve the system to be very reliable, and to incorporate some learning strategy. We then included CBR to help provide the historical trend in data as well as to perform the vital search for a new case solution and best therapy.

This initial effort used CBR for diagnoses validation aided by what we call *Case Similarity Index* (*CSI*). The *CSI* is a value generated from patient's physical features, namely: age, sex, and body weight as well as the symptoms slot numbers in that design. The three main physical features, provided the cases in the case base with a broad *categorization* in terms of context, while the symptoms factors helped deal with low level differentiation needed between cases within a category.

To accomplish this task, we designed the use of a novel method that assigns relevant *case weights* to all patient cases stored in the case base. Our approach uses the *Singular Value Decomposition* (*SVD)* method related to *principal component analysis* in combination with the *MVF* above as detailed in (Ochi-Okorie, 1997).

In the algorithm, each new case diagnosed uses the disease class and the case weight as seed to the CBR unit in order to retrieve best five to ten matching cases. (The 5-10 figure was chosen because of the limit of 10 cases per disease category and the available viewing window on the custom GUI – *Graphical User Interface*). Refinements and selection of a winning case was then done by comparing current patient's complaint (description text) with the best of the 5-10 top cases from the case base search.

The 10 case limit design constraint in the case base is intended to reduce storage requirements as well as improve our search times. For example, the case base will always have no more than 10 best cases of *Malaria*, 10 best cases of *Typhoid Fever*, 10 best cases of *Filariasis*, etc. As cases were added, a pattern began to emerge as related cases in disease classes (such as Malaria, or Typhoid Fever) began to form *clusters* (distinct groups) based on the *case weights.* Internally, our case base indexing and retrieval depended heavily on the case weights since they were generated from actual patient case specific features. Our approach uses the class means and covariance values to weed out any cases from the case base whenever we exceed 10 in a given class. The learning from this approach, and the details of the *SVD* techniques are beyond the scoop of this paper. However, suffice it to say that we realized great improvements in our diagnoses to nearly 98%.

With the CBR unit built and containing a good number of cases, we decided to make the individual units to perform the diagnoses independently, and thereafter "*vote*" for the wining disease class. The *voting* mechanism simply collects all the diagnosis from each unit, and selects the most popular disease class. This *CBR integration* scheme is illustrated in *Figure 1* and discussed further in the next section.

We would however still rely on the CBR unit for retrieving appropriate therapy actions for the wining disease class because of its ***memory*** capabilities or advantages over the other methods of reaching a diagnosis. For this major reason, this architecture that is said to be *CBR dominant* (Reategui and Campbell, 1994) in which the control is biased more toward the CBR component in the integration. It provides the solution (therapy) based on the selected diagnosis class and assists in the learning algorithm from the associated case weights.

In their work, Reategui and Campbell suggest four possible ways of integrating CBR with other approaches in reasoning. These are *central control, distributed control, dominant, and non-dominant control,* respectively.

3.1 The Voting Approach

The new *voting* approach described here is still experimental, but uses some of our earlier techniques such as the *MVF* for pre-processing data. Essentially, four distinct components (shown in *Figure 1*) capable of performing independent diagnosis combine their results in order to filter out the best result for the whole system.

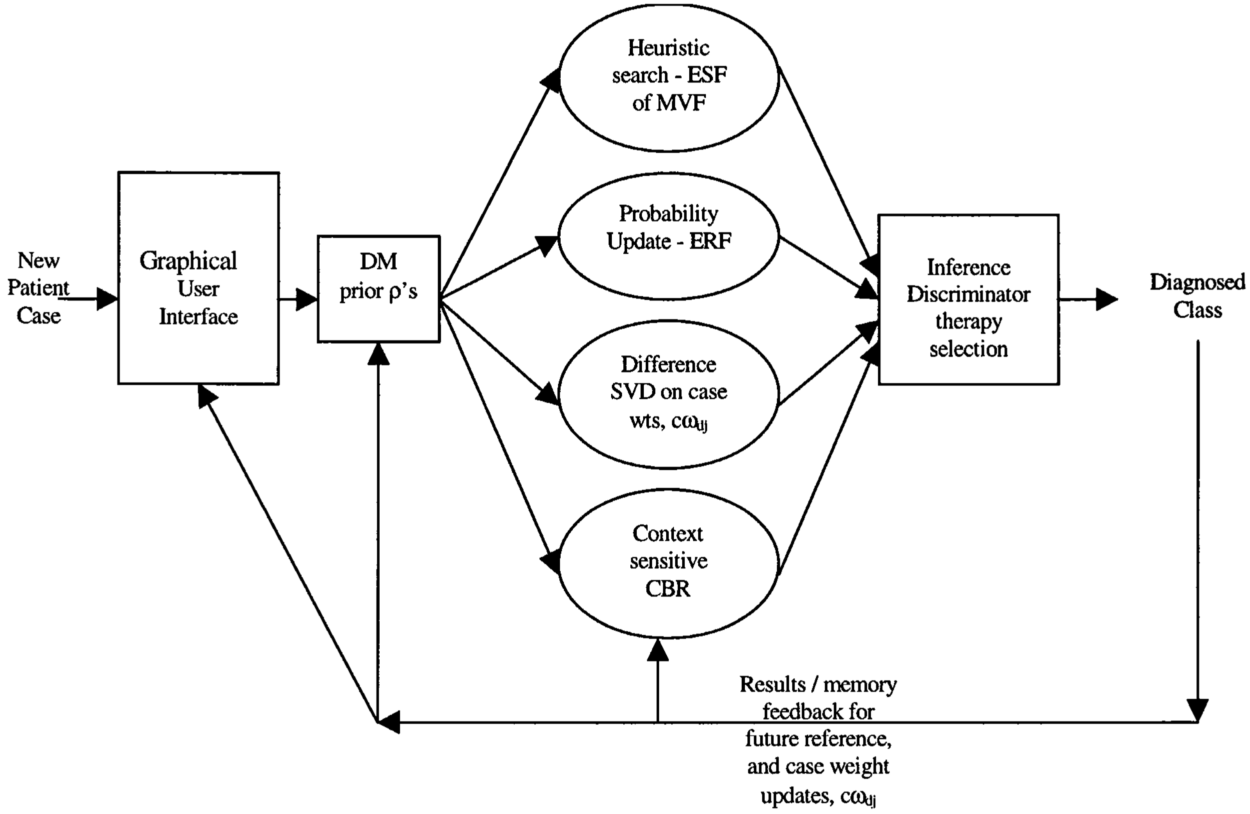
New Patient Case
Graphical User Interface
DM prior ρ's
Heuristic search - ESF of MVF
Probability Update - ERF
Difference SVD on case wts, cωij
Context sensitive CBR
Inference Discriminator therapy selection
Diagnosed Class
Results / memory feedback for future reference, and case weight updates, cωij

Figure 1: Diagnoses by Voting in CBR-Dominant Integration

The *MVF* - our *Extended Similarity Functions*, uses the DM (*Disease decision Matrix*) which a binary encoded table of symbolic domain knowledge (ideal expert knowledge). The DM is then used to find a solution with its algorithmic functions by heuristically matching it with the transformed (binary values) new case data. Let us now suppose it found disease d_1 as the culprit.

The *ERF* with probability updates, uses a learning algorithm to compute new *ERF's* and compare them with ideal *ERF* initially specified from expert knowledge and the a priori probabilities for disease-symptom associations. New values continuously update or replace initial a priori values as data is gathered in the on-going effort. It looks at all symptoms, and pre-disposing factors (pdf's) of each disease class as associated features and analyzes how they have occurred in actual clinical data collected on new cases. (The *ERF* learning algorithm is not covered here for space constraints.) In this way, it is possible to compare expert's initial best guess to see how consistent it is with actual values for a given geographic area. Pdf's may vary widely for different areas. For example, pdf's associated with Malaria in Northern Nigeria does not include swampy rain forest as in most parts of the South or Niger Delta. Thus, we have a suggested d_2 based on best ratio of evidence and the revised probabilities.

Using the initial DM data from the *MVF inference engine* the *SVD* computes *case weights* for all possible classes in relation to the new case features. The *SVD* computes case weights based on the detected dominant disease class for each new case feature set. The best case match is one with an *SVD difference* of approximately zero when compared with the *ideal SVD* value for each disease class. Now let us suppose the wining case class is d_3. Wide disparities between the two values indicate a wider difference between the new case and the ideal case class for each of the tropical diseases in the study.

The CBR unit performs a search based on the given new case complaints as "*Search Description*" text. It also uses the externally supplied "*Case Similarity Index*" (*CSI*) which is a numerical value generated from symbolic case features in combination with three high level *context* features - age, sex, and body weight of the patient. Each case stored has its own associated therapy plan and reference to patient's personal records / medical history. Further, let us suppose the CBR unit's diagnosis suggests d_4 as the winner.

3.2 The Voting Algorithm

Our *voting algorithm* will then collect all four diagnoses (d_1, d_2, d_3, and d_4) and compare them (see the Appendix). A success count is taken for two or more identical diagnoses. If for example, there are more than one diagnoses of Malaria in the d_j set above, then our big culprit is *Malaria.* If we have a tie, such as two diagnoses of *Malaria*, and two diagnoses of *Typhoid Fever*, then we know we have a clear case of ***differential diagnosis*** in which the actual therapy would exist in the region between the therapies for both diseases for the new patient case.

To "force" or bias the system into avoiding a tie, we make the CBR unit to *dominate* the decision making process by requesting a special case base search using the *SVD* case weights. The CBR unit on being flagged to revise it's result will perform extra statistical analyses on the case base data for the two tied disease classes using the stored case weights for all the cases in those classes. The mean and covariance case weight values of the two contending classes are then compared with that of the new case. The *winning* class is the class with case weight closest to the new case weight and the *CSI* feature value. Thus, the difference between these covariance and mean values would be approximately zero for the best match.

As with the CBR - ANN paradigm, the CBR dominance can be seen more clearly here for ties in diagnoses in addition to the fact that search results / therapies retrieved from the case base constitute the final action ordered for the ill patient.

In the very rare event that all four units arrive at different diagnoses, our procedure will follow a similar "forced" path as outlined above for finding a wining class. This situation although not yet encountered is a design check for when it does arise. Thus, in this manner our new approach provides better diagnostic results than what was the case in our previous work. The figure is approximately 98% accuracy compared with 86-90% earlier.

4.0 Results and Conclusion

By adding the case weights generated with SVD, some interesting patterns in clinical data became rather obvious. Average case weights for 3 to 10 cases in each concluded diagnosis in the training data set of the case base were very distinct from one class to the other. For example: Malaria was 0.2626, Typhoid Fever 0.1334, Amoebiasis 0.0187, Cholera 0.0293, Shigellosis (Bacillary Dysentery) 0.0151, Tuberculosis 0.0479, Tetanus (LockJaw) 0.0335, etc.

By using it as an additional index, case base searches were quicker and more efficient since they were stored pre-ordered according to their indexes. The case base retrievals were the best 5 using these case weights. Generally, since good retained cases were the only ones retrieved, we almost always had good results from past clinical cases. For cases where prognosis was "Patient died", "death", "expired" and the likes, we forced the case base to reject those cases no matter how good their weights were. Further, using the case weights in addition to patient-case ID, and CSI, a snap shot of the case base always showed some potentially useful data clustering (aggregations). Introduction of the *difference SVD* for heuristic rule learning is particularly useful in improving the results. The SVD algorithm in concert with the CBR unit perform learning from cumulative knowledge in the case weights which we have used as search index in place of the description text or features in patients complaint. The new logic in **TROPIX** uses the difference between the average covariance values from existing cases and the new case features expressed in the same dimensionless weights.

In the CBR Integration, the final selection of a wining case is done by a voting algorithm, which takes the most popular diagnosis from all the collaborating reasoning modules as peers, thus providing higher accuracy. The model then selects the best therapy based on the wining disease category, or class as the final diagnosis and hence, the therapy for a new patient case.

It has been recognized that we need to account for errors in the data arising from SME acquisitions or clinical records in order to minimize their effects on the overall system performance of the multi-modal logic integrating CBR. The storage capabilities of the CBR element, and the periodic update by various statistical methods in our processing will hopefully help in cleaning out any "fuzzy" data in the system.

References

1. Joly, S. and Le Calve', G., "*Similarity Functions*" in Van Custem, B. (Ed.), *Classification and Dissimilarity Analysis*, LN in Statistics, Springer-Verlag, New York, 1994.
2. Kock, G. *The neural network description language CONNECT, and its C++ implementation*, Technical report, GMD FIRST Berlin, Universitat Politecnica de Catalunya, August 1996.
3. Krovvidy, S. and Wee, W.C. *Wastewater treatment systems for case-based reasoning*. Machine Learning, 10:341,1993.
4. Lees B. and Corchando, J., *Case Based Reasoning in a Hybrid Agent-Oriented System*, 5th German Workshop on Case-Based Reasoning, March 1997.
5. Lees B., Rees, N. and Aiken, J. *Knowledge-based oceanographic data analysis,* Procs. Expersys-92, Eds. F. Attia, ET. Al, IITT Int'l Paris, October 1992, pp. 561-65.
6. Ochi-Okorie, A. S., *Disease diagnosis validation in TROPIX using CBR*, Artificial Intelligence in Medicine Journal, Elsevier Science Publishers, Amsterdam, The Netherlands, 12/1, pp. 43-60 January 1998.
7. Ochi-Okorie, A.S., *Combining Medical Records with Case-Based Reasoning in a Mixed Paradigm Design - TROPIX Architecture & Implementation*, LN in Artificial Intelligence, CBR R&D - Second Int'l Conf. On CBR, ICCBR-97, Providence, RI, July 1997, Proc.; editors David B. Leake and Enric Plaza; Springer Verlag.
8. Reategui, E. and Campbell, J. A. *A classification system for credit card transactions.* Procs. Second European Workshop on Case-Based Reasoning, pp. 167-174, November 1994.

Appendix

1. **Integration name/category**: A *CBR dominant multi-strategy design* with *voting* collaboration
2. **Performance Task**: Medical or related diagnosis / decision support, especially for differential diagnosis or second opinion generation.
3. **Integration Objective**: Facilitate basic health care delivery with improved diagnostic accuracy. Accuracy is determined by comparing actual clinical cases diagnosed by system with the diagnostic results provided by clinicians on the same cases. Obtain differential diagnoses in cases belonging to different tropical disease classes but with close similarities in symptoms, clinical signs, and patients' *pre-disposing factors* (*pdfs*).
4. **Reasoning Components**: CBR - for *context sensitive* (high level categorization of cases) search and indexing of cases using the seeded values of the *Case Similarity Index* (*CSI*), and the more granular case weights generated by the *Singular Value Decomposition* (*SVD*) operation to refine search results from case memory.

ERF - Evidence Ratio Factors which simply computes the ratio of observed symptoms / signs and *pdfs* in the cases against their expected counts for a disease class. The mean of these values for each disease class is used to update expert projected a priori probabilities for the symptoms / signs / pdfs.

ESF of *MVF - The Extended Similarity Functions* (*ESF*), a part of the *Matched Vector Functions* (*MVF*) that basically perform *heuristic* matching between the new case features and the ideal (or prototype case) / disease class. It looks for similarities, dissimilarities, and for missing features in the compared cases.

SVD difference - Evaluates the *difference* between the prototype case weights and the new case weight. The closer the value to zero, the better the match. ($\mathit{SVD\_class\text{-}wt} - \mathit{SVD\_case\text{-}wt} \cong 0$)

5. **Control Architecture**: CBR dominance (partial Master-Slave) in that it over-rules when a tie exists in the votes from all the reasoning components. It does so by performing a new search / retrieval of cases based on revised case weights using statistical means and covariances.
6. **CBR Cycle Step(s) Supported**: Retention / Storage, Retrieval, Reuse, and Revision.
7. **Representations**: Patient cases as vector objects with multiple features, and features as numeric / symbolic attributes of diseases in a patient. Associated patients' personal (static data) data, case therapies and pre-existing conditions of patients as medical records in a relational data base table.
8. **Additional Components**: Statistical processing / pattern matching, and database clustering to improve indexing and retrieval of records.

9. **Integration Status**: Applied experimentally building upon previous work in **TROPIX**.

10. **Priority future work**: Learning algorithms, clinical evaluation of results, and feature text (natural language description of patient complaints) translation to numeric search indexes.