Experiences in exploiting data at the Universitat de Girona: Selected problems in application domains.

Joaquim Meléndez, Joan Colomer, Beatriz López, Josep Vehí,

Carles Pous, Magda L. Ruiz, Luis E. Mújica

Institut d'Informàtica i Aplicacions, Universitat de Girona,
Campus Montilivi, Edifici P.4. Girona, Spain.

Contact:
quimmel@eia.udg.es

Abstract: Experiences in heterogeneous application domains treated with different data mining approaches are presented in this paper: Case based reasoning and self organising maps have used to diagnose beams and pipes after analysing their responses using wavelet decomposition. Also case based reasoning methodology has been used to improve electronic circuits diagnosis based on selected exemplars retained according to data base reduction polices. Waste water treatment plants have been monitored using different statistical models based on extensions of principal component analysis. Intelligent agents are used in web based domains to obtain a unique model for users interactuating in multiples contexts. Motivation and main results are highlighted in the papers.

1 Introduction

Nowadays, amounts of data are available in many domains where quality control, traceability and/or monitoring are obligated by law. In other domains the simply necessity to better know system capabilities or to improve functionality or knowledge provokes the existence of important data repositories. The actual necessity is to develop methodological solutions to automatically extract significant information from these data useful enough to built decision models.

This goal motivated the impulsion of a transversal action of three research groups[1] of the Universitat de Girona on promoting data mining activity to cope with specific problems as forecasting, diagnosis or automatic profiling. This text summarises some representative experiences where data mining techniques were proposed as solving methodologies. Next section presents two diagnosis applications using a Case Based Reasoning (CBR) approach. Statistical methods involving Principal Component

[1] For additional information of research activity of these groups: **eXiT (http://exit.udg.es), ARLAB (http://eia.udg.es/arl), MICE (http://mice.udg.es)**.

Analysis (PCA) are used in section 3 to model a waste water treatment plant in ordere to discover abnormal behaviours. Section 4 reviews a data integration problem in web based domains under a agent point of view to build a unique user models useful to operate in multiple domains. Last section concludes the main contributions of these experiences.

2 Case based reasoning for diagnosing industrial systems

Diagnosis in some domains (civil structures, diagnosis of electronic devices or many medical diagnosis based on pattern recognition) is only possible after measuring system response (dynamics) and extracting relevant features, or attributes, from these signals in order to be associated with known diagnostics. Several methods can be applied to perform this abstraction procedure (wavelet decomposition, segmentation, discretisation, frequency analysis, polynomial regression, etc.) suitable for diagnosis. Figure 3 shows how wavelet transform is applied with this proposal in a case-example discussed below.

Two examples are presented in the following subsections in which has been made use of the existence of a mathematical model to generate enough data to implement a case based diagnosis system. In the first example a differential equations model has been used to simulate an analogue circuit whereas in the second one finite element simulation has applied to characterise faults in pipes and beams. Case Based Reasoning (CBR) methodology has been implemented in both examples allowing learning capabilities and diagnosis based on instances generated from these models. Case base organisation and maintenance have a decisive role in retrieving appropriate cases for diagnosis as the following examples show.

2.1 Exemple 1: Electronic circuit diagnosis

The process of testing or diagnosing circuits consists in applying certain types of excitations to a circuit and then analyzing the responses obtained in order to derive a possible failure. Dictionary based methods are commonly used for diagnosing circuits in a two steps procedure: fault signatures gathering to build the dictionary, and a matching to identify a new fault according to the dictionary. A typical dictionary is based on exciting the circuit with a saturated-ramp waveform to obtain four signature parameters: The steady-state (V_{est}), the overshoot (SP), the rising time (t_r) and the delay time (t_d). Drawbacks appear in such systems when considering tolerance effects giving poor and wrong diagnosis in a high per cent of situations. Storing more cases seems to be a solution to improve the percentage of diagnosis successes but it has been demonstrated that spoils the dictionary performance due to overlapped measures obtained due to tolerance of components. Hence, there is a compromise between fault coverage and dictionary length.

Reduction techniques have been applied to a Monte-Carlo generated case base of faulty signatures. Dictionary size reduction has then applied using instance pruning techniques, such as IB3, DROP4 and All-KNN [9]. These techniques are based on the

improvements of the nearest neighbour algorithm. System performance is improved giving learning capabilities according to CBR cycle to this new dictionary [10].

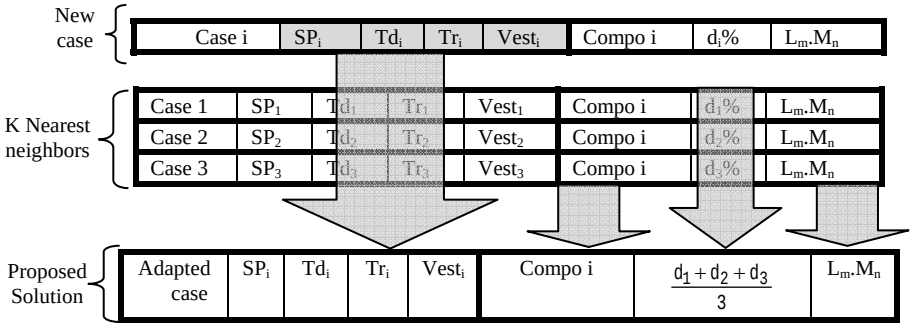

Figure 1. Case adaptation

. Fault dictionary definition has been extended to cases adding structural information described as a hierarchical decomposition of the circuit representing faults by its location, the component and its exact deviation from nominal value (expected). A metric function and a k-NN (Nearest Neighbour) retrieval function have been defined. Figure 1 exemplifies a diagnosis performed after retrieving 3-NN previously diagnosed.

Once the solution (diagnosis) to the new case (faulty circuit) is proposed, the CBR cycle is completed with a revision of this case in order to be or not retained. If the solution is considered correct and accurate enough, it is not necessary to retain the new case. On the other hand, if it is considered to be incorrect or with poor accuracy, the new case will be retained in the case memory. But, previously to its retention, it is analyzed if this new case retention is going to spoil other cases already contained in the case base classification. This process is done using DROP4 algorithm. The revision analyzes how the cases that constitute the adapted solution are performing the diagnosis. When the CBR-system is testing circuits with unknown faults, there is no revision task, since the proposed diagnosis can not be contrasted with the correct one.

	Classic	Spread	DROP4	IB3
Predicted faults (±20% and ±50%) with tolerances	82.04%	80.68%	82.36%	78.64%
Non predicted faults (±70%)	17%	16.2%	17.625%	17.5%
Case base size	25	12500	1112	2457

Table 1. Dictionaries results for previously predicted and non predicted faults

Also a maintenance police, based on IB3, is implemented in the system to avoid the *utility problem* factor. The same criterion as IB3 is used for cases removing, that is, when the performance of a particular case drops below a certain established value with a certain confidence index, the case is considered to be spoiling the diagnosis and it will be deleted. The confidence limits used in IB3 are the ones defined by the success probability of a Bernoulli process.

Table 1 depicts the percentage of diagnosis success obtained for a biquadratic filter (a benchmark). The first row shows how the different dictionaries perform for a

set of 2500 predicted faults corresponding to deviations of ±20% and ±50% from the nominal value for each component, but taking into account the tolerance effect in the others. In spite of having considerably more cases, observe that the spread dictionary has a lower diagnosis success compared to the classic and reduced by DROP4 dictionaries. The second row demonstrates how the diagnosis drops drastically for 2500 non previously considered situations, that is, any deviation compressed in the range of ±70% for each of the components while the others stay into their nominal range. The third row shows the dictionary size.

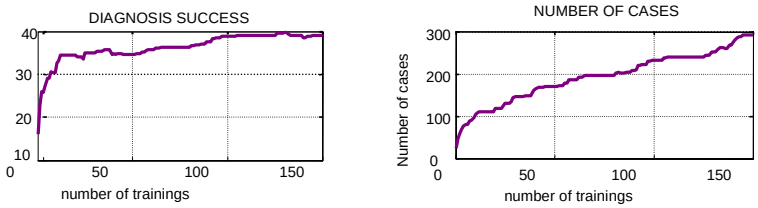

Figure 2. Evolution of diagnosis success while training.

Results obtained using the proposed CBR-system increases considerably the percentage of diagnosis success, while keeping a moderate case base size. Figure 2 shows the progress of training related to success ratio. The training is stopped at nº 153, where there is a high percentage of success and a reasonable case base size. In particular, for the set of non previously predicted faults with deviations compressed in the range of ±70% from nominal, the diagnosis success is 39.375% while the case base size after the training is 263.

2.2 Example 2: Structural assessment of pipes and beams

The problem of damage identification in structural analysis is usually studied under the phenomenon of elastic strain wave propagation. An excitation signal is applied and the resulting dynamic response is examined. A DM approach based on the integration of Case-Based Reasoning –CBR-, Self Organizing Maps -SOM- (as a classification tool in order to organize the old cases in memory) and Wavelet Transform -WT (to extract features from the measured signal) has been developed [7].

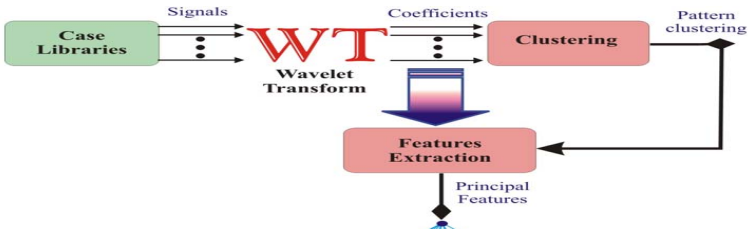

Figure 3 Example of feature extraction for case representation

Cases are available from both simulated structures and real system under test in order to be used for diagnosis by analogy. Features are extracted from coefficients of the Wavelet Transform applied to dynamic response. Wavelet coefficients are used to represent cases and clustered according to a distribution function (Figure 3). The case

base is then reorganised using a Self Organizing Map (SOM) in order to organize in each output neuron or cluster the cases with similar characteristics ([7]) as Figure 4 shows. The unified distance matrix of this SOM (U-matrix) (Figure 5a) indicates the distances among weights of each neuron and its neighbourhood. High values involve small correlation between clusters. The distribution of a variable into the SOM before and after training it is shown in Figure 5b and Figure 5c.

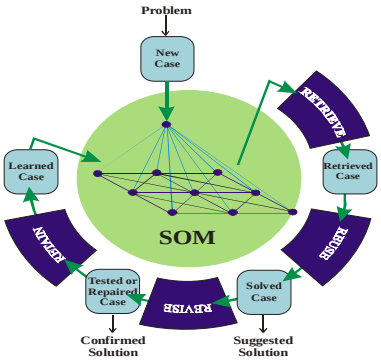

Figure 4 Modified CBR cycle

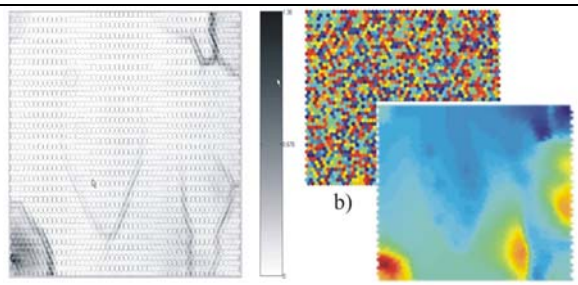

Figure 5 Self Organizing Map. a) U-matrix.
b) Disorganized variable before learning
c) Organized variable after learning

Finally, diagnosis is performed according to a statistical criterion of retrieved cases. Size and intensity of defects to be diagnosed is adapted from retrieved cases by a weighted function of them. Weights depend on distances (histogram distribution) of the retrieved cases. This methodology has been applied to two types of structures: beams (see Figure 6a) and a pipes (see Figure 6b) where materials and geometric specifications have previously been determined to perform an accurate generation of the case base.

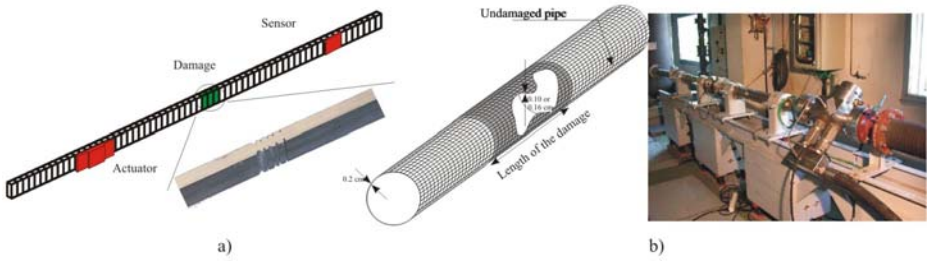

Figure 6 a) Beam. b) Pipe

Beam diagnosis is performed dividing in 92 elements and defects can be present in lengths of maximum 5 elements. 5464 cases of damaged structure have been simulated (up 10 defective consecutive elements with 12 different mass reductions). It spent around 10 hours. 57 principal features have been extracted from each signal. A SOM of 57 input neurons and 50*50 output neurons has been trained in 35 minutes. As test example a damage on elements -44-45-46- (see Figure 6a), is detected and

diagnosed applying the methodology identifying he fault in the elements -46- with a reduction of mass of 45%.

For the pipes example, they have been divided in 16 sections and defects in each section by reducing its thickness in 20% and 50% around the pipe, have been simulated. After producing a on the third element with a thickness reduction of 20%, the system locates the damaged element with a mass reduction of 31.72%.

3 Statistical models for Wastewater Treatment Plant (WWTP) monitoring

Waste water treatment plant (WWTP) are difficult to be automatically monitored because the behavioural dependence with external factors, as the quality of influent or weather (temperature, rain), that affects biological activity of microorganisms. Within the aim of exploiting a new type of WWTP (Sequencing Batch Reactor, SBR) and an automatic monitoring system, a statistical model based on Multiway Principal Component Analysis (MPCA). This SBR-WWTP (Figure 7) is operated in fixed duration cycles or batches (8 hours in the pilot plant; See Figure 8) where anoxic (without oxygen, when denitrification occurs) and aerobic (with oxygen, when nitrification occurs) conditions are alternated to fulfil the complete operation. After the whole cycle sludge containing nitrogen and organic matter is eliminated and cleaned water drawn. The measured variables (pH, Oxidation Reduction Potential (ORP), Dissolved Oxygen (DO) and Temperature) are sampled every 5 seconds and registered resulting in a 5760 samples array per variable. The goal is to automatically assess the plant behaviour from the analysis of these data.

Figure 7 SBR pilot plant and Schematic

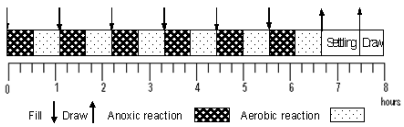

Figure 8 Fixed operational cycle applied in the pilot plant

3.1 Batch process statistical modelling with MPCA and Multiblock MPCA

Multivariate Statistic Process Control (MSPC) methods are commonly used to detect and diagnose variation in dynamic systems produced by external agents (instrument failures, human intervention, weather,…) by analysing correlations between variables and building lower dimension models. Fundamentals of MSPC applied to

batch processes are extensions of Principal Component Analysis (PCA) and partial least squares (PLS) to multivariate systems (MPCA,MPLS). In this work only extensions of PCA have been used [1] under the following assumptions:

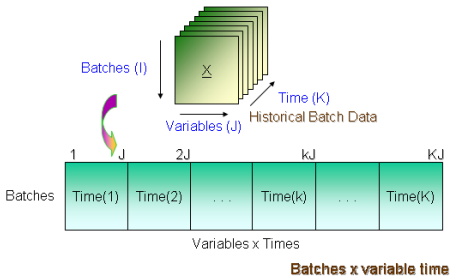

Figure 9 Decomposition of X to 2-D (I x KJ)

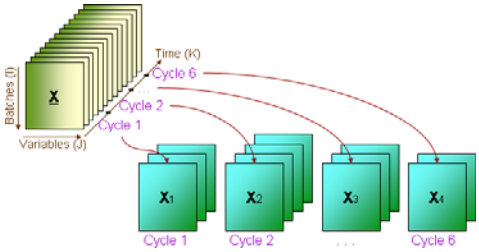

Figure 10 Dividing the 3-D matrix into different stages

Consider a cycle (or batch) run in which j = 1, 2, ..., J variables are measured at k = 1, 2, ...,K time instants throughout the batch. Similar data will exist on a number of such batch runs i = 1, 2, ..., I. The X (*I* x *J* x *K*) array collects data for a batch. This 3-way array (X) can be decomposed into a large 2-dimensional matrix by unfolfding data in the batch direction (a *I* x *KJ* matrix is then obtained) as it is depicted in Figure 9. Thus, Principal Component Analysis ([2]) can be applied to obtain a statistical model in a lower dimension space by correlation analysis. After the analysis a set of ordered Principal Components are obtained as a linear combination of previous data.

Multiblock MPCA is another extension that divides the data matrix (*I* x *KJ*) into *K* blocks (X_1, X_2,…, X_K) grouping all the samples gathered in the same instant in a unique block (Figure 10). This approach has significant benefits because the latent variable structure is allowed to change at each phase in the batch processes. Each data block is considered as a separate source of information and used to obtain its own model [3]. Control charts based on Q-statistic and T^2 are commonly used for batch process monitoring. Q_i indicates the distance between the actual values of the batch and the projected values onto the reduced space whereas T^2 gives a measure of the Mahalanobis distance in the reduced space between actual batch and the average model previously obtained [4].

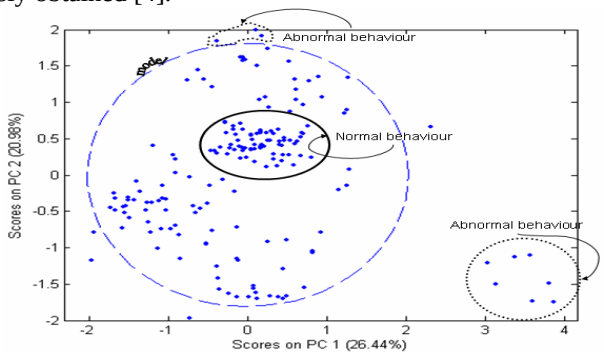

Figure 11 Score plot for batches

This methodology has been tested with a set of 179 available batches resampled and scaled to normalise its influence in the statistical model. A 179 x 4 x 392 array,

X, has been unfolded in the batch direction resulting a (179 x 1568) array. PCA discovers 8 principal components explaining the 92.79% of the total variability. In Figure 11 a projection on the first and second component plane of the statistical model (dashed line) and the 179 batches is given. Five typologies of batch have been determined (Table 2): electrical fault, variation in the composition, equipment defects, atmospheric changes (raining) and normal behaviour.

Abnormal behaviour	Quantity	%
Atmospheric Changes	17	9,50
Equipment defects	8	4,47
Variation in the composition	33	18,44
Electrical Fault	2	1,12
total	60	33,52
Normal behaviour	Quantity	%
Excellent	98	54,75
Good	14	7,82
Normal	7	3,91
total	119	66,48

Table 2 Quantity and percentage of batches for each type

Q

Abnormal behaviour	Quantity	%
Atmospheric Changes	9	5,03
Equipment defects	0	0,00
Variation in the composition	11	6,15
Electrical Fault	0	0,00
total	20	11,17
Normal behaviour	Quantity	%
Excellent	5	2,79
Good	3	1,68
Normal	0	0,00
total	8	4,47

T^2

Abnormal behaviour	Quantity	%
Atmospheric Changes	4	2,23
Equipment defects	6	3,35
Variation in the composition	8	4,47
Electrical Fault	2	1,12
total	20	11,17
Normal behaviour	Quantity	%
Excellent	0	0,00
Good	0	0,00
Normal	0	0,00
total	0	0,00

Table 3 MPCA classification according to Q and T^2.

The combined use of Q and T^2 charts allows a first analysis based on the whole batch behaviour. Table 3 summarises this results: 31 about 60 of the abnormal behaviour can be detected, 9 batches are in both charts. T^2analysis allows to detect two batches related with electrical fault (EF). A fine analysis can be consulted in [5]. The application of Multblock MPCA (Figure 10) allows to a more accurate analysis by splitting batches into stages (a batch consist of 9 stages as is depicted in Figure 8). In such a case the use o 7 principal components offers 93% of representativity. Q and T2 analysis is used to fire alarms at every stage allowing a diagnosis of variation in composition in batches 11 to17. Combination of MPCA and Multiblock MPCA offers a complete statistical analysis of process behaviour.

4 Agent-based web mining for user integration information

The evolution of Internet has enabled to citizens to access to an ever expanding amount of information in which the user hardly find the information he/she is looking for. This is known as the overload information problem for which several techniques have proved useful in helping users to handle the large amount of information. Mainly, personalisation and web mining techniques has been widely accepted. On one hand, personalisation techniques keep information on user interests and data in order to make systems provide the appropriate answer to the user at the right moment and place. In order to achieve personalisation, most system relies on either user profiles or user models On the other hand web mining concerns the use of data mining techniques to automatically discover and extract information from web documents and services [11]. There are several web mining categories, among which web usage mining is the one in which the UdG group has been working along several years.

Web usage mining is the type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers. Thus, web usage mining allows the discovery of user profiles and user models than can then be used to personalize system interactions.

In this section two applications are presented that illustrate both, personalisation and web usage mining. First, a case-based reasoning system is described that learns user profiles in a recommendation process. And second, an integration information framework is presented that allows user information shift from one domain to the other. Both approaches are deployed on open distributed environments and agent-based systems are used as the supporting technology.

4.1 Case-based reasoning for recommender systems

Recommender systems use personalisation techniques to help users to locate particular items, information sources and people that best match their interests or preferences [12]. In order to achieve such goals, recommender systems relies on machine learning techniques that build user profiles. In this section we focus in the use of case-based reasoning (CBR) to model the user in which cases capture both explicit interest (the user is asked for information) and implicit interest (captured from the user interaction) of a user on a given item. Then, CBR is used to recommend items to the user according to the similarity of past user interest kept on the case base. Such process is known as content-base filtering,.

CBR is a well known paradigm on the machine learning community that the UdG group has been traditionally working on as the previous sections has shown. However, when CBR is applied to recommender systems, several drawbacks arise. Mainly, the approach addressed for case-base maintenance. That is, the uncontrolled growth of the number of past experiences should be tackled not due to the data case coverage as traditionally has been performed, but taking into account the adaptation of the system to the users' changing interest over time.

In order to deal with such adaptation issue, a drift attribute has been assigned to each case. Such attribute controls the age and relevance of the case. Then, a forgetting mechanism updates the drift attribute according to the user-system interaction. Details on both, the drift attribute and the forgetting mechanism can be found in [13].

Content-based filtering implemented by means of CBR is one of the possible procedures to follow in order to recommend items to a user. However, there are other approaches known as collaborative filtering in which the recommendations of the systems are sharply improved due to the inclusion of information about other users. Such collaborative method, however, requires the revelation of personal information about the user. In order to maintain the privacy of the users' personal data, intelligent agents can be used. Then, a social model of users involved in a collaboration environment has been deployed, in which each user is represented by an intelligent agent (see Figure 12). Each agents keeps user interest according to the CBR methodology together with a trust model of its neighbours that relates the affinity of the user interest with other agents in the community. The trust value with which agents label their neighbours is obtained following a playing agents procedure and updated according

to a trust adaptation method, both explained in detail in [14]. Thanks to the trust model, collaboration is achieved while keeping user data on privacy.

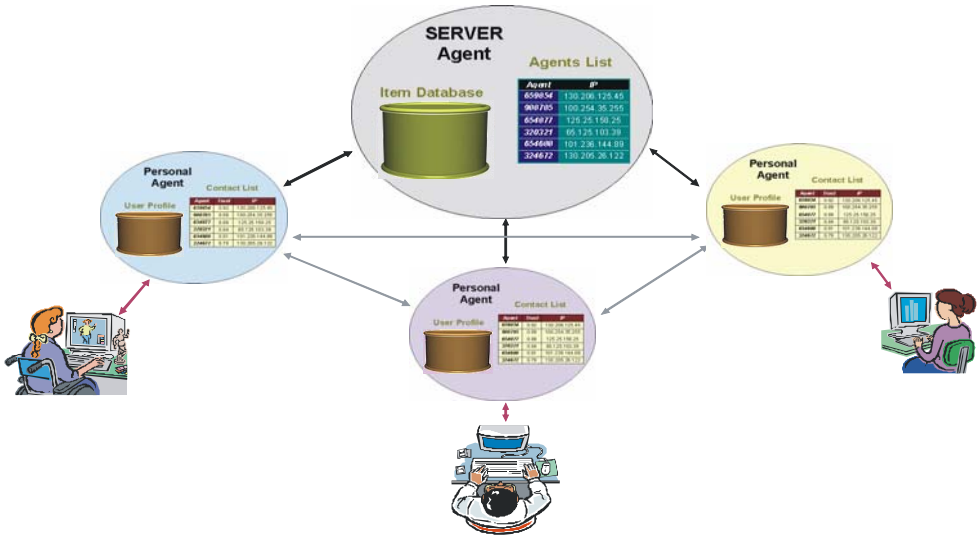

Figure 12. Collaborative CBR agents.

The system implemented with the methodology described here is GenialChef, that has been awarded with the Special Prize to the best system actually deployed in the Agentcities network and with the Prize to the Best University Project of E-Tech 2003.

4.2 Smart user models for multiple domains

The current scenario in recommender systems is given by the interaction of one user model to one recommender system. This means that the user has several user models according to the number of applications which he/she interacts (see Figure 13). In this scenario, the user must provide his/her information whenever he/she needs a service. In addition, the user models do not share a common structure and vocabulary about the user across applications. These limitations do not allow the possibility of sharing the user model in different domains.

The aim of Smart User Models is to keep the information related to a user in an integrated way. In [15] and [16] a Smart User Model methodology from a multiagent perspective is presented. Three representation levels of the user have been defined: the cognitive level, the computational level and the domain level. Such user model structure facilitatea the user data transfer, from one domain, in which the user has already been profiled, to another, with which the user has never before interacted.

The key issue in this research is the definition of similarity measures that allow the identification of common concepts from one domain to the other. Most of the similarity measures depend on weights that should be automatically learned. Shannon information measures are being explored as a possible alternative.

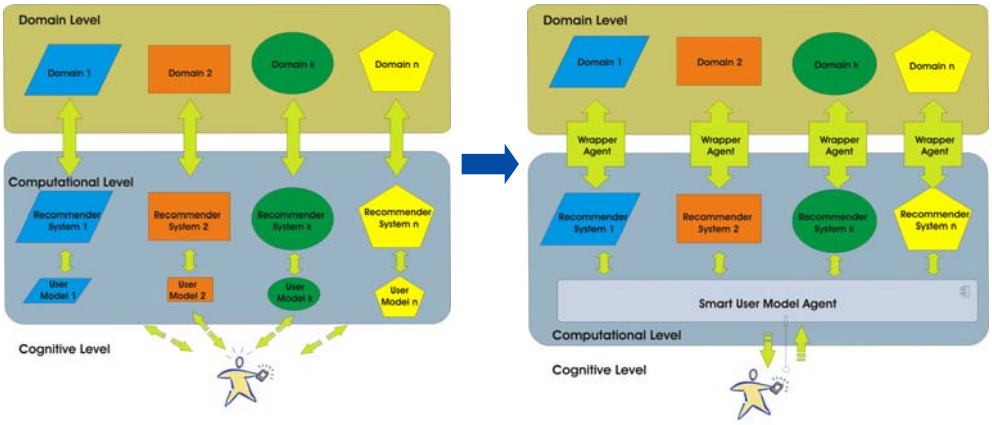

Figure 13. Left: Current user models. Right: smarts user models.

5 Conclusions

Several data mining techniques have been explored to cope with specific problems. Dependence of solving technique with domain and goals has been made evident. Nevertheless some interesting conclusions can highlighted from each problem.

Feasibility to assess structures using case based-reasoning has been demonstrated through numerical and experimental tests. The main value or innovation is the exploitation of the model of the structure to pre-load the case base and the improving by constant feedback through real structure. Similar integration of models and instance based reasoning have been implemented to diagnose electronic circuits. In this problems the used of adequate maintenance police increases diagnosis performance and reduces data base making it computationally efficient.

Multivariate Statistical Process Control preserves accuracy (of 92% of the initial information in the WWTP problem) at same time that dimension is reduced. MSPC allows relating the batch behaviour with types of batch process defined previously and dividing the process into meaningful blocks it is possible to localize better the batches with abnormal behaviour, allowing for clear indication of the types of batch process like normal behaviour, atmospheric changes, etc.

Finally, agent technology has shown a convenient platform to combine several learning techniques (as case-based reasoning) in order to share and integrate information involving user interaction on the Internet.

6 Acknowledgements

This work is a summary of results from several research projects funded by MCYT from the Spanish Government and European Union (FEDER funds) under the following contracts: SECSE-DPI2001-2198, SBR-WT-DPI2002-04579-C02-01, and DPI2001-2094-C03-01.

References

[1] Theodora Kourti, “Process analysis and abnormal situation detection: From theory to practice”, IEEE Control Systems Magazine, 22(5), 10–25, (oct 2002).

[2] Paul Nomikos and John F. MacGregor, “Monitoring batch processes using multiway principal component analysis”, AIChE, 40(8), 1361– 1375, (aug 1994).
[3] S. Joe Qin, Sergio Valle, and Michael J. Piovoso. “On unifying multiblock analysis with application to decentralized process monitoring”. Journal of Chemometrics, (15):715–742, 2001.
[4] Evan L. Russell, Leo H. Chiang, and Richard D. Braatz. “Data Driven Techniques for Fault Detection and Diagnosis in Chemical Processes”Advances in Industrial Control”. ISBN 1-85233-258-1, London, 2000.
[5] Ruiz M, Colomer J., Rubio M., and Meléndez J. “Combination of multivariate statistical process control and classification tool for situation assessment applied to a Sequencing Batch Reactor wastewater treatment” ISCQ 2004 ISBN 83-88311-69-7 pp: 257-267 print House Zaklad Poligraficzny Jerzy Kosinski, Warszawa
[6] Melendez J, Colomer J, de la Rosa JL, (2001), “Expert Supervision Based on Cases”, Proceedings of the 8th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA’0 1, pp: 431-440.
[7] Mujica LE, Vehí J, Rodellar J, García O, Kolakowski P,(2004), “Hybrid Knowledge Based Reasoning Approach for Structural Assessment”, Proceedings of the Second European Workshop on Structural Health Monitoring, pp 591-599.
[8] Balivada A., Chen J., Abraham J.A. (1996), “Analog Testing with Time Response Parameters”, IEEE Design and Test of Computers, pp 18-25.
[9] Wilson D., Martinez T. (2000). “Reduction Techniques for Instance-Based Learning Algorithms”, Machine Learning Vol 38 (3), pp 257-286.
[10] Pous C., Colomer J., Melendez J. (2004), “Extending a Fault Dictionary Towards a Case Based Reasoning System for Linear Electronic Analog Circuits Diagnosis”. To be published on the Proceedings of the European Conference on Case Based Reasoning (ECCBR04).
[11] Kolari, P., Joshi A. “Web Minint: Research and Practice”. IEEE Computing in Science and Engineering, July/August 2004.
[12] Montaner M., López, B., de la Rosa, J.Ll. A Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19: 285-330, June, 2003.
[13] Montaner, M., López, B., de la Rosa, J. Ll. Improving Case Representation and Case-Based Maintenance in Recommender Agents. Lecture Notes in Computer Science (Artificial Intelligence) 2416:234-248, 2002. ECCBR’02.
[14] Montaner, M., López, B., de la Rosa, J. Ll. Opinion-based Filtering through Trust. Lecture Notes in Computer Science (Artificial Intelligence) 2446: 164-178, 2002. CIA’02.
[15] González, G., López B. and de la Rosa, J. LL. Smart User Models for Tourism: A Holistic Approach for Personalised Tourism Services. Information Technology & Tourism Journal, Volume 6. Num. 4. 2004.
[16] González, G., López B. and de la Rosa, J. LL. Managing Emotions in Smart User Models for Recommender Systems. In Proceedings of 6th International Conference on Enterprise Information Systems ICEIS 2004. Volume. 5. pp. 187-194. April 14-17, 2004.