Enfoque UTE, V.7-Sup.1, Feb.2017, pp.232 - 243
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542 / p-ISSN: 1390-9363

Recibido (Received): 2017/01/09
Aceptado (Accepted): 2017/02/24
CC BY-NC-ND 3.0

Razonamiento basado en casos aplicado al diagnóstico médico utilizando clasificadores multi-clase: Un estudio preliminar

(Case based reasoning applied to medical diagnosis using multi-class classifier: A preliminary study)

D. Viveros-Melo [1], M. Ortega-Adarme [2], X. Blanco Valencia [3], A. E. Castro-Ospina [4]
S. Murillo Rendón [5], D. H. Peluffo-Ordóñez [6]

Abstract:
Case-based reasoning (CBR) is a process used for computer processing that tries to mimic the behavior of a human expert in making decisions regarding a subject and learn from the experience of past cases. CBR has demonstrated to be appropriate for working with unstructured domains data or difficult knowledge acquisition situations, such as medical diagnosis, where it is possible to identify diseases such as: cancer diagnosis, epilepsy prediction and appendicitis diagnosis. Some of the trends that may be developed for CBR in the health science are oriented to reduce the number of features in highly dimensional data. An important contribution may be the estimation of probabilities of belonging to each class for new cases. In this paper, in order to adequately represent the database and to avoid the inconveniences caused by the high dimensionality, noise and redundancy, a number of algorithms are used in the preprocessing stage for performing both variable selection and dimension reduction procedures. Also, a comparison of the performance of some representative multi-class classifiers is carried out to identify the most effective one to include within a CBR scheme. Particularly, four classification techniques and two reduction techniques are employed to make a comparative study of multi-class classifiers on CBR.

Keywords: Case based reasoning; High dimensionality; Variable selection.

Resumen:
CBR ha demostrado ser apropiado para trabajar con datos de dominios poco estructurados o situaciones donde es difícil la adquisición de conocimiento, como es el caso del diagnóstico médico, donde es posible identificar enfermedades como: cáncer, predicción de epilepsia y diagnóstico de apendicitis. Algunas de las tendencias que se pueden desarrollar para CBR en la ciencia de la salud están orientadas a reducir el número de características en datos de gran dimensión. Una contribución importante puede ser la estimación de probabilidades de pertenencia a cada clase para los nuevos casos. Con el fin de representar adecuadamente la base de datos y evitar los inconvenientes causados por la alta dimensión, ruido y redundancia de los mimos, en este trabajo, se utiliza varios algoritmos en la etapa de pre-procesamiento para realizar una selección de variables y reducción de dimensiones. Además, se realiza una comparación del rendimiento de algunos clasificadores multi-clase representativos para identificar el más eficaz e incluirlo en un esquema CBR. En particular, se emplean cuatro técnicas de clasificación y dos técnicas de reducción para hacer un estudio comparativo de clasificadores multi-clase sobre CBR.

Palabras clave: Razonamiento basado en casos; alta dimensión; selección de variables.

[1,2] Universidad de Nariño, Pasto – Colombia (dianavive.77@udenar.edu.co, mabel12-02@udenar.edu.co)
[3] Universidad de Salamanca, Salamanca – España (xiopepa@usal.es)
[4] Tecnológico Metropolitano, Medellín – Colombia (andrescastro@itm.edu.co)
[5] Universidad Autónoma de Manizales, Manizales – Colombia (smurillo@autonoma.edu.co)
[6] Universidad Técnica del Norte, Ibarra – Ecuador (dhpeluffo@utn.edu.ec)

1. Introduction

Case-based Reasoning (CBR) solves new problems by retrieving previously solved problems and reusing the corresponding solutions. The specificity of the case-based approach of reasoning lies in its focus on the inseparability of reasoning from memory and from learning (Bichindaritz & Conlon, 1996) In CBR terminology, a case usually denotes a problem situation. A previously experienced situation, which has been captured and learned in a way that it can be reused in the solving of future problems, is referred to as a past case, previous case, stored case, or retained case. Correspondingly, a new case or unsolved case is the description of a new problem to be solved. Case-based reasoning is - in effect - a cyclic and integrated process of solving a problem, learning from this experience, solving a new problem, etc. (Aamodt & Plaza, 1994).

The common life cycle for solving a problem using CBR is mainly carried out in four-step process (Trendowicz & Jeffery, 2014):

1. Retrieve: One or more problems that are similar to the new problem are retrieves from the base of previously solved problems, and one attempts to modify them to fit the new problem parameters.
2. Reuse: The solutions of the selected previous problems are reused to solve the new one.
3. Revise: The solved new problem is then revised against the actual solution.
4. Retain: When successfully tested, it is added to the base of previous problems to be reused for solving future problems.

In particular, CBR has demonstrated to be an appropriate methodology for:

- Working with unstructured domains data or difficult knowledge acquisition situation, for example, formal models or universally applicable guidelines do not well understand many diseases (Herrero, 2007), (Bichindaritz & Marling, 2006).

- Help Desk systems used in the area of customer service to solve problems with products or services (Jenal, Gonzales, Alejo, & Ramos López, 2006).

- Predictions of the possible success of a proposed solution when the information is stored considering the level of success of the solutions, the case-based reasoned can be able to predict proposed solution to the current problem. Clearly, the reasoner will have in not only those levels of success stored but also the differences between the Case or cases recovered and the current situation (Lozano & Fernández, 2008).
- Automatic Acquisition of Subjective Knowledge because CBR systems exhibit an incremental knowledge acquisition, and knowledge can be abstracted by generalizing cases (Phuong, Hoang, Prasad, Hung, & Drake, 2001).

- Medical diagnosis based on the similarity of the symptoms of the base of cases with the current case select the one that best suits to make a diagnosis (Jenal, Gonzales, Alejo, & Ramos López, 2006).

Among these tasks, medical diagnosis has been one of the most popular research subjects in both medical informatics and computer science communities. Medical diagnosis is the process aiming at identifying diseases based on findings (such as symptoms, lab reports, patient's complaints, and other environmental factors) Adapting computer aided decision support systems to the diagnosis process requires a database-like medical knowledge base, and a problem-solving strategy applied to the knowledge base. A single problem solving strategy may be sufficient for simple diagnosis problems However, some difficulties may arise when the diagnosis problem becomes complex (Wang, Hsien-Tseng, & Tansel, 2013).

So, the CBR, is a reasoning process, which is medically accepted and also getting increasing attention from the medical domain. A number of benefits of applying CBR in the medical domain have already been identified (Bichindaritz & Marling, 2006), (Gierl, Bull, & Schmidt, 1998), (Montani, 2008). However, the medical applications offer several challenges for the CBR researchers and drive advances in research (Begum, Ahmed, Funk, Xiong, & Folke, 2011).

In order to adequately represent data and to avoid the inconveniences caused by its high dimensionality, we propose the use of variable selection and dimension reduction techniques in a preprocessing stage for CBR tasks, finally, we make a comparative study of multi-class classifiers to assess processed data performance.

The rest of this paper is structured as follows: Section II describes the proposed methodology, as well as the pattern recognition procedures used in this work. Section III presents the proposed experimental setup. Results and discussion are gathered in section IV. Finally, some concluding remarks and future works are drawn in Section V.

2. Material and Methods

This section outlines the proposed framework to assess the feasibility of using multi-class schemes within CBR approaches. Particularly, we resort to the adaptation of a pattern recognition stages into the CBR life cycle.

In the CBR scheme, the recovery is the most important stage, since in this phase the system finds the most similar cases to the current unknown case, simulating an efficient memory as a human expert would (Kolodner, 1983). By combining the CBR methodology with classifiers, a cost function would be used to find the nearby cases.

The next stage where we adapt classifiers would be in the adaptation stage, because we want to show the answer in terms of probabilities. With the classifier, we can find the membership degree of the new case in each of the classes, which would be helpful for medical staff.

To that end, we propose to carry out a comparative study of multi-class classifiers within preprocessing, recovery, and adaptation CBR stages. Fig. 1 depicts the proposed methodology to perform the comparison of multi-class classifiers.

A. *Preprocessing*

Variable selection: First, as preprocessing stage a variable selection procedure is employed. In this work, we use the so-called correlation based feature subset (CfsSubsetEval) algorithm, which evaluates the relevance of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy among them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. And as search method the bestfirst algorithm, to reduce the number of parameters per instance of a dataset with a backtracking. It starts with the whole set of attributes and search backward to reduce the number of parameters per instance of a dataset.

Dimensionality Reduction: After performing variable selection and aiming to improve both visual inspection and classification performance, a dimensionality reduction stage is employed by using well known methods, namely Laplacian Eigenmaps (LE), that uses spectral techniques to perform dimensionality reduction. This technique relies on the basic assumption that the data lies in a low-dimensional manifold in a high-dimensional space (Belkin, 2003) and t-distributed stochastic neighbor embedding (t-SNE), it is a nonlinear dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot.

B. *Adaptation and recovery*

Here, with the aim of accomplishing a multi-class case recovery, representative multi-class classifiers are considered. Due to their characteristics, we select the following classifiers: *K* Nearest Neighbor Classifier (*K*-NN) being a geometric-distance-based-approach, artificial neural networks (ANN) being a heuristic-search-based approach, support vector machines (SVM) being a model-based classifier, and Parzen's Classifier (PC) being a non-parametric density-based classifier.

3. Experimental Setup

A. *Database*

For evaluating the proposed methodology, we used two databases from UCI Machine Learning Repository. The first one, named Cardiotocograms, contains 2126 fetal cardiotocograms belonging to different classes. The dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians. This data set consists of 21 attributes which include LB - FHR baseline (beats per minute), AC of accelerations per second, FM of fetal movements per second, UC of uterine contractions per second, DL of light decelerations per second, DS of severe decelerations per second, DP of prolonged decelerations per second, ASTV percentage of time with abnormal short term variability, MSTV mean value of short term variability, ALTV percentage of time with abnormal long term variability, MLTV mean value of long term variability, Width width of FHR histogram, Min minimum of FHR histogram, Max Maximum of FHR histogram, Nmax of histogram peaks, Nzeros of histogram zeros, Mode - histogram mode, Mean histogram mean, Median histogram median, Variance histogram variance, Tendency histogram tendency, CLASS FHR pattern class code (1 to 10) and NSP fetal state class code (Normal=1; Suspect=2; Pathologic=3).

The second database, named Cleveland, contains 303 instances. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). Consisting of 13 attributes which include age, sex, chest pain type (Typical angina=1; Atypical angina= 2; Non-anginal pain=3; Asymptomatic=4), resting blood pressure, cholesterol, fasting blood sugar (True=1; False=0), resting electrocardiographic results (Normal=1; Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)=2; Showing probable or definite left ventricular hypertrophy by Estes' criteria=3), maximum heart rate, exercise induced angina (Yes=1; No=0), oldpeak, slope (Upsloping=1; Flat=2; Downsloping=3), number of vessels coloured, thal (Normal=3; Fixed defect=6; Reversible defect=7) and the classification values from 0 no presence to 4 types of heart diseases.

B. Parameter settings and procedures

As outcomes of the preprocessing stage, we obtain that Cardiotocograms database is reduced to 10 features, and Cleveland database to 7 features. Subsequently, as part of the same stage, by using dimensionality reduction techniques Cardiotocogram database is reduced to a 2-, 3-, 5-, 8-dimensional space. Likewise, Cleveland database is reduced to 2-, 3-, 5-dimensional space. As well, the whole subset of selected variables is considered for both databases.

For classification techniques, it should be stated out that a 20-fold cross-validation was performed to achieve unbiased results. Particularly, the following setup is established:

- *K-NN:* Is a nonparametric supervised classification method based on distances. This instance-based classification technique needs a value for the number of neighbors (K), such parameter is optimized by means of a leave-one-out strategy, is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the approximate function is trained on all the data except for one point and a prediction is made for that point.

- *ANN*: The heuristic-based classification technique requires a number of units per hidden layer. In this work, a back-propagation trained feed-forward neural net is used with a single hidden layer. The number of units is computed from the data itself as the half of the instances divided by feature size plus the number of classes. The weight initialization consists of setting all weights to be zero, as well as the dataset is used as a tuning set.

- *SVM*: This instance-based classification method takes advantage of the kernel trick to compute the most discriminative non-linear hyperplane between classes. Therefore, its performance heavily depends on the selection and tuning of the kernel type. For this work a Gaussian kernel is selected given its ability of generalization and its band-width parameter was fixed by the Silverman's rule (Sheather, 2004).

- *PC*: This probabilistic-based classification method requires a smoothing parameter for the Gaussian distribution computation, which is optimized.

As a performance measure, it is used the standard mean classification error.

4. Results and Discussion

Achieved results for different number of dimensions as well as different classifiers are shown in Table I as the mean and standard deviation over the 20 folds runs. It can be seen how Cleveland dataset is a challenge task since performance is poor for all classifiers. It should be stated also that dimensionality reduction does not necessarily improves classification performance for both dimensionality reduction techniques. Nevertheless, by reducing dimensionality there is a gain in visual analysis of data as can be appreciated in *Figure 1*, particularly it can be seen how in 2D (*Figures 2(a) and 2(c)) and 3D (Figures 2(b) and 2(d*)) Cleveland data is highly overlapped which is consistent with achieved results. It should be noted that the error for SVM classifier is 0.397 ± 0.07, which is not far from the result obtained in (Bhatia, Praveen , & Pillai, 2008), where the classification accuracy with 7 attributes is of 70.36%.

Table 1. Achieved classification performance over 20-fold cross validation for considered databases and dimensionality reduction techniques

DB	Reduction Technique	# dimd	*K*-NN	ANN	SVM	PC
Cleveland	t-SNE	2	0.381 ± 0.08	0.389 ± 0.067	0.389 ± 0.013	0.393 ± 0.093
		3	0.382 ± 0.06	0.367 ± 0.09	0.389 ± 0.013	0.393 ± 0.069
		5	0.397 ± 0.07	0.362 ± 0.089	0.389 ± 0.028	0.4 ± 0.087
		7	0.397 ± 0.07	0.347 ± 0.062	0.401 ± 0.029	0.393 ± 0.069
	LE	2	0.408 ± 0.069	0.393 ± 0.077	0.389 ± 0.013	0.393 ± 0.041
		3	0.397 ± 0.066	0.397 ± 0.075	0.389 ± 0.013	0.374 ± 0.047
		5	0.389 ± 0.067	0.404 ± 0.085	0.412 ± 0.036	0.389 ± 0.067
		7	0.389 ± 0.065	0.382 ± 0.065	0.397 ± 0.07	0.404 ± 0.064
Cardiotocograms	t-SNE	2	0.037 ± 0.015	0.084 ± 0.038	0.071 ± 0.017	0.077 ± 0.017
		3	0.036 ± 0.016	0.073 ± 0.02	0.054 ± 0.019	0.076 ± 0.018
		5	0.032 ± 0.017	0.088 ± 0.017	0.039 ± 0.019	0.075 ± 0.016
		8	0.035 ± 0.016	0.079 ± 0.016	0.033 ± 0.017	0.075 ± 0.019
		10	0.031 ± 0.017	0.082 ± 0.036	0.028 ± 0.016	0.076 ± 0.019
	LE	2	0.045 ± 0.014	0.078 ± 0.016	0.086 ± 0.017	0.102 ± 0.023
		3	0.054 ± 0.018	0.072 ± 0.015	0.061 ± 0.016	0.09 ± 0.02
		5	0.042 ± 0.014	0.075 ± 0.031	0.048 ± 0.014	0.09 ± 0.016
		8	0.039 ± 0.015	0.067 ± 0.019	0.038 ± 0.013	0.065 ± 0.016
		10	0.381 ± 0.25	0.06 ± 0.017	0.038 ± 0.016	0.063 ± 0.016

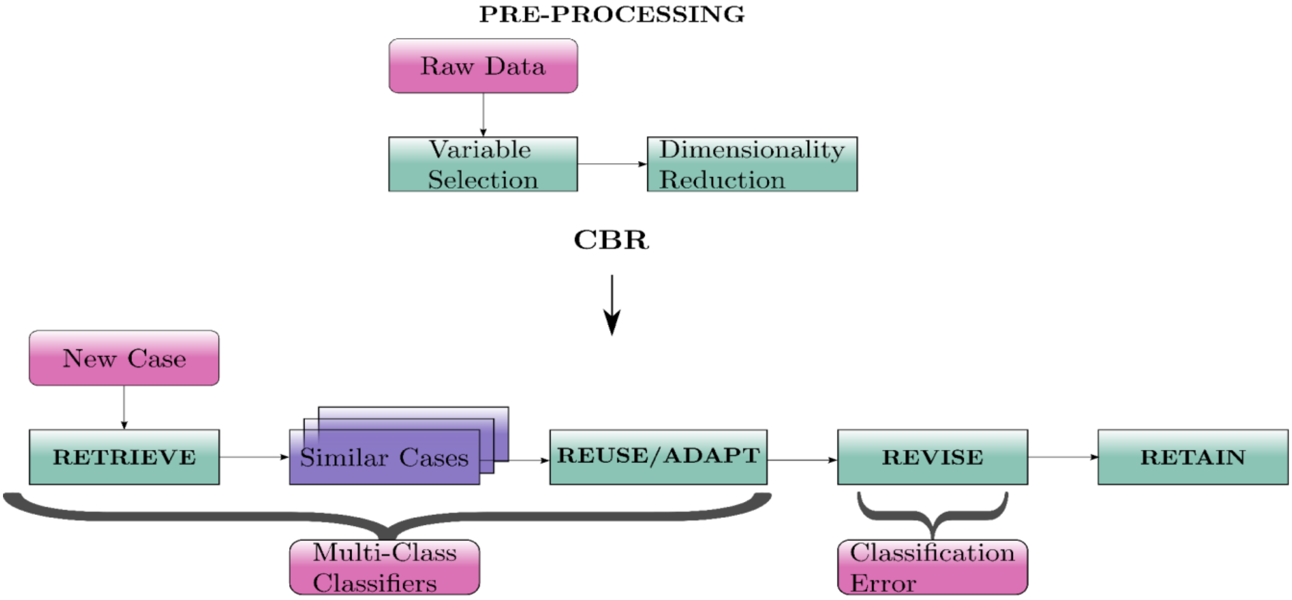
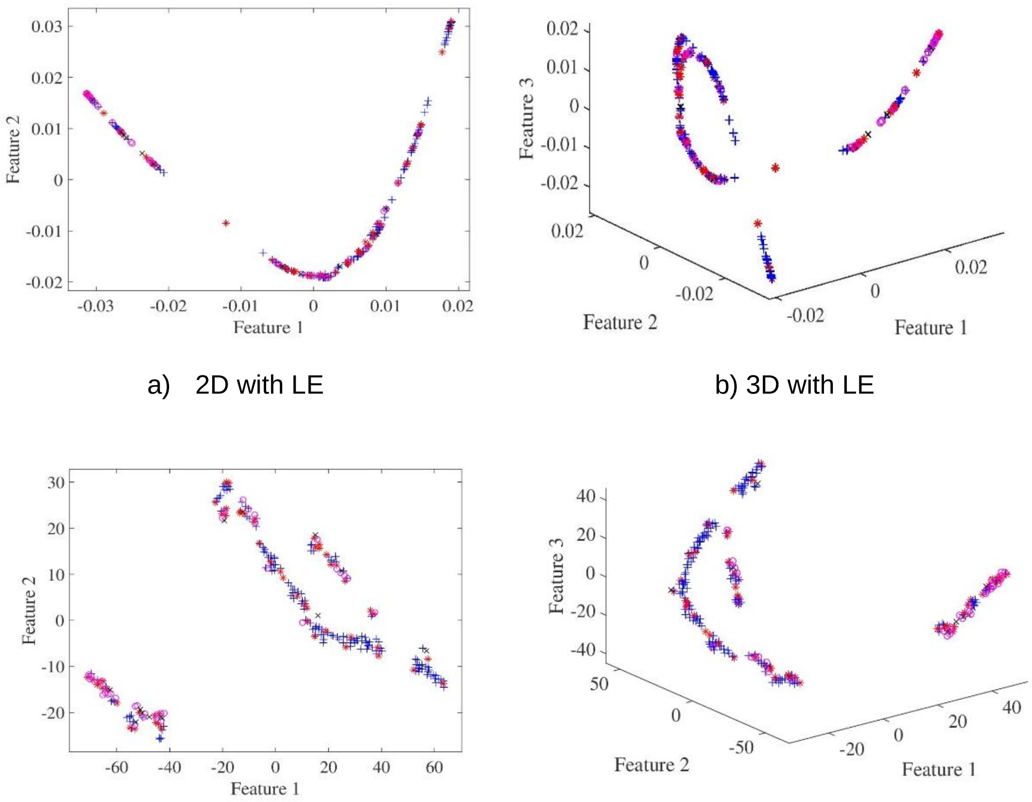


Figure 1. Block diagram of proposed methodology. The aim of the comparative study is assessing the possibility of incorporating multi-class classifiers into CRB approaches design, as well as identifying the best classifier for this task



a) 2D with LE

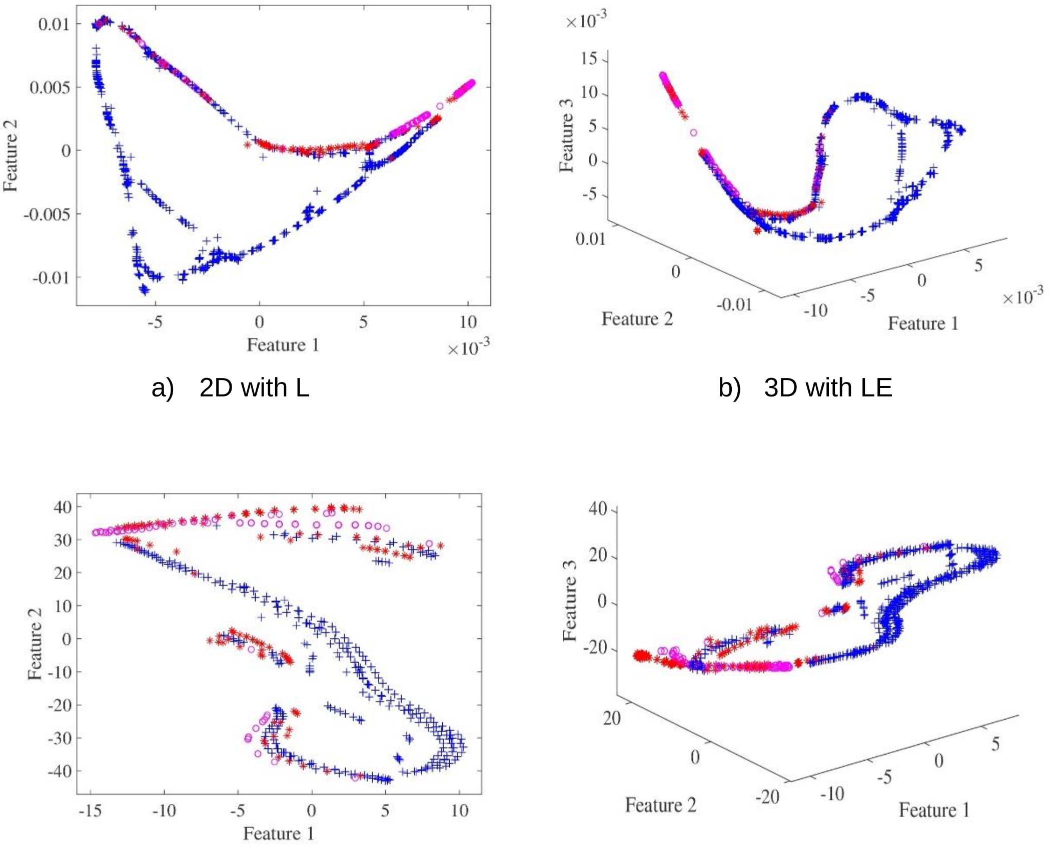
b) 3D with LE

c) 2D with t-SNE

d) 3D with t-SNE

Figure 2. Low-dimensional scatterplots for Cleveland database. Figures (a), (c) show the first two features from database. Figures (b), (d) show the first three features from database.

For cardiotocograms dataset classes separability is evident in lower dimensions, i.e. 2D and 3D, as depicted in Figures 3(a) to 3(d) leading to outstanding results as shown in Table I, however, as for Cleveland dataset, dimensionality reduction does not substantially improve classification performance on Cardiotocograms dataset even though it enhances data visualization. We can see that for the Cardiotocograms database the best result was using the SVM classifier the error is 0.028 ± 0.016, improving the results obtained in (Sundar, Chitradevi, & G. , 2012) where they achieved an average accuracy of 0.9328.



a) 2D with L

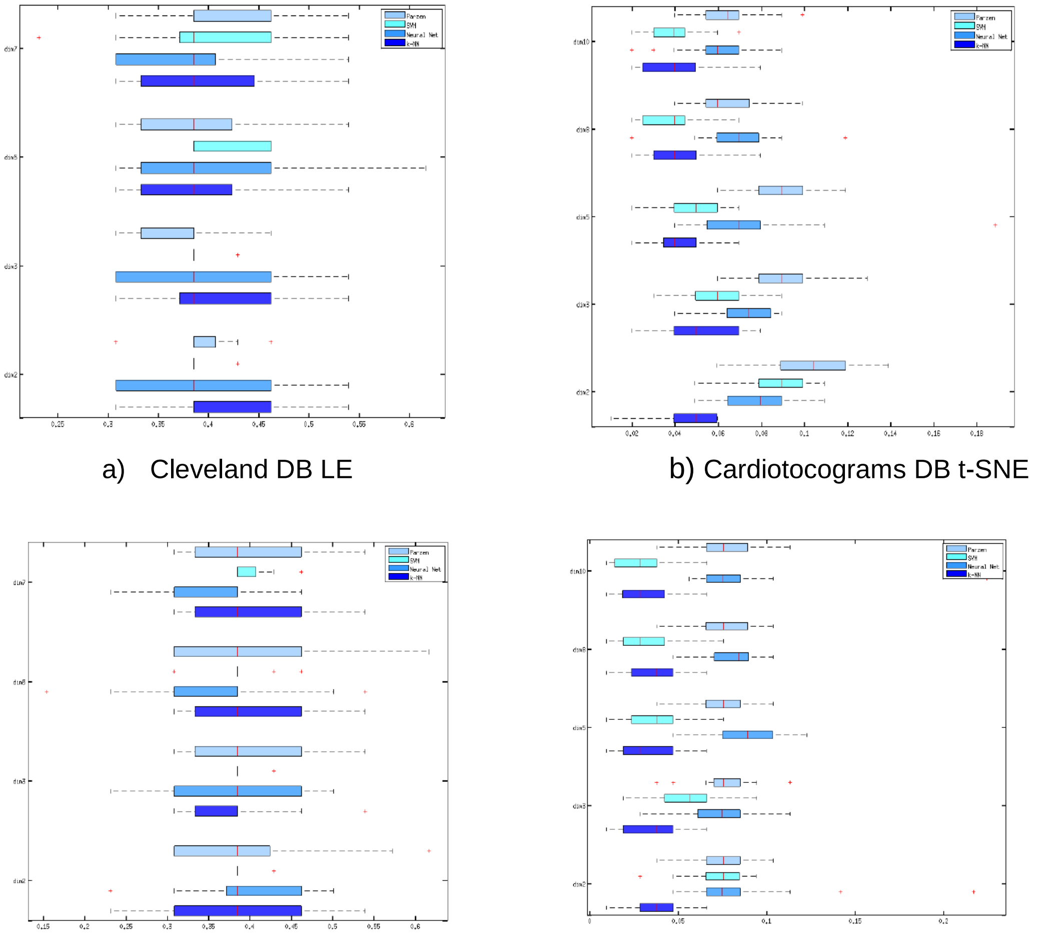
b) 3D with LE

c) 2D with t-SNE

d) 3D with t-SNE

Figure 3. Low-dimensional scatterplots for Cardiotocograms database. Figures (a), (c) show the first two features from database. Figures (b), (d) show the first three features from database.

By performing a stability assessment, it could be seen from Figure 4 by the width of the error boxplots how SVM and *K*-NN classifiers achieves the best results for considered Cardiotocogram and Cleveland databases. Moreover, it should be noted how SVM classification results are the most stable of the considered classification techniques.



a) Cleveland DB LE

b) Cardiotocograms DB t-SNE

c) Cleveland DB t-SNE

d) Cardiotocograms DB t-SNE

Figure 4. Classification error boxplots for considered classification techniques on Cleveland and Cardiotocograms databases

5. Conclusions and future work

This work presents a feasibility evaluation of the use of techniques from the field of pattern recognition into CBR frameworks, so that conventional CBR can be extended to multi-class scenarios. The above, with the aim of facilitating decision making in the diagnostic support, especially in situations where subcategories may exist and even emerging categories according to the condition of the patients.

Experimentally we prove that the SVM classifier is a good candidate for integration with the CBR approach to create a generic system to assist physicians in the diagnosis of patients and is capable of working with databases multiclass associating probabilities each class, responding to one of the challenges of (Bichindaritz & Marling, 2006), (Kwiatkowska & Atkins, 2004).

As a future work, we will explore the possibility to design a case recovery stage for CBR able to deal with multi-class cases while providing users with class membership (probabilities to belong) estimates for a new case, improving usability with respect to conventional approaches and, in addition, providing more meaningful information to the expert at the review stage in order to provide a more accurate diagnosis.

To improve the performance of the classifiers, we will study new techniques of variable selection and reduction of dimensions, as well as techniques of data balancing for databases that require it as is the case of Cleveland, due to problems such as class overlapping or the lack of data representative of input data.

Acknowledgments

Authors would like to thank to the Facultad de Ingeniería en Ciencias Aplicadas as well as electronic engineering and telecommunications program from Universidad Técnica del Norte.

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications 7*, 39-59.

Begum, S., Ahmed, M. U., Funk, P., Xiong, N., & Folke, M. (2011). Case-based reasoning systems in the health sciences: a survey of recent trends and developments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 421-434.

Belkin, M. (2003). Problems of learning on manifolds. *The University of Chicago*.

Bhatia, S., Praveen , P., & Pillai, G. (2008). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. En *Proceedings of the World Congress on Engineering and Computer Science, WCECS* (págs. 22-24).

Bichindaritz, I. &. (1996). Temporal knowledge representation and organization for case-based reasoning.

Bichindaritz, I., & Conlon, E. (1996). Temporal knowledge representation and organization for case-based reasoning. *Temporal Representation and Reasoning, 1996.(TIME'96), Proceedings., Third International Workshop on*, 152-159.

Bichindaritz, I., & Marling, C. (2006). Case-based reasoning in the health sciences: What's next? *Artificial intelligence in medicine*, 127-135.

Gierl, L., Bull, M., & Schmidt, R. (1998). CBR in Medicine. En *Case-Based Reasoning Technology* (págs. 273-297). pringer Berlin Heidelberg.

Herrero, J. M. (2007). *Una aproximación multimodal al diagnostico temporal mediante razonamiento basado en casos y razonamiento basado en modelos. Aplicaciones en la medicina.*

Jenal, Gonzales, M., Alejo, S. M., & Ramos López, R. (2006). Sistema CBR para presentación de entrenamientos físicos personalizados en Internet.

Kolodner, J. (1983). Maintaining organization in a dynamic long-term memory. *Cognitive science 7*, 243-280.

Kwiatkowska, M., & Atkins, M. (2004). Case representation and retrieval in the diagnosis and treatment of obstructive sleep apnea: a semio-fuzzy approach. En *Proceedings of the 7th European Conference on Case-Based Reasoning* (págs. 5-35).

Lozano, L., & Fernández, J. (2008). Razonamiento basado en casos: Una visión general.

Montani, S. (2008). Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. *Applied Intelligence 28*, 275-285.

Phuong, Hoang, N., Prasad, N., Hung, D. H., & Drake, J. (2001). pproach to combining case based reasoning with rule based reasoning for lung disease diagnosis. En *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th* (págs. 883-888). IEEE.

Sheather, S. (2004). Density estimation. *Statistical Science 19*, 588-597.

Sundar, C., Chitradevi, M., & G. , G. (2012). Classification of cardiotocogram data using neural network based machine learning technique. *International Journal of Computer Applications 47*.

Trendowicz, A., & Jeffery, R. (2014). *Software project effort estimation: Foundations and best practice guidelines for success.* Springer.

Wang, Hsien-Tseng, & Tansel, A. U. (2013). MedCase: a template medical case store for case-based reasoning in medical decision support. En *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (págs. 962-967). ACM.