A multi-class extension for case-based reasoning applied to medical problems: A first approach

D. Viveros-Melo*
M. Ortega-Adarme+
Universidad de Nariño
Pasto, Colombia
Email: *dianavive.77@udenar.edu.co
+mabel12-02@udenar.edu.co

X. Blanco Valencia
BISITE Research Group
Universidad de Salamanca, Spain
Email: xiopepa@usal.es

A. E. Castro-Ospina
Research Center of the Instituto
Tecnológico Metropolitano
Medellín, Colombia
Email: andrescastro@itm.edu.co

S. Murillo Rendón
Universidad Autónoma de Manizales
Manizales, Colombia
Email: smurillo@autonoma.edu.co

D. H. Peluffo-Ordóñez
Universidad Técnica del Norte
Ibarra, Ecuador
E-mail: dhpeluffo@utn.edu.ec

***Abstract*—Case-based reasoning (CBR) is a problem solving approach that uses past experience to tackle current problems. CBR has demonstrated to be appropriate for working with unstructured domains data or difficult knowledge acquisition situations, as it is the case of the diagnosis of many diseases. Some of the trends and opportunities that may be developed for CBR in the health science are oriented to reduce the number of features in highly dimensional data, as well as another important focus on how CBR can associate probabilities and statistics with its results by taking into account the concurrence of several ailments. In this paper, in order to adequately represent the database and to avoid the inconveniences caused by the high dimensionality, a number of algorithms are used in the preprocessing stage for performing both variable selection and dimension reduction procedures. Subsequently, we make a comparative study of multi-class classifiers. 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.

I. INTRODUCTION

Case-based Reasoning (CBR) solves new problems by retrieving previously solved problems and reusing the corresponding solutions. In the past twenty years, CBR methodology has attracted much attention, showing its usability in applications usually focused on open and weak theory domains, such as medical diagnosis, design, corporate planning and many engineering domains [1]. The core of the CBR is the case, which usually indicates a problem situation. From another point of view, a case is prior learning experience, which has been captured and can be reused to solve future problems. The life cycle for solving a problem using CBR is mainly carried out in four phases: to identify the current problem and find a past case similar to the new case (retrieve), using the case and suggest a solution to the current problem (reuse/adaptation), evaluate the proposed solution (revise), and update the system to learn from experience (retain) [2].

The CBR has demonstrated to be an appropriate methodology for:

- Working with unstructured domains data or difficult knowledge acquisition situation, for example, many diseases are not well understood by formal models or universally applicable guidelines [3], [4].
- Making tasks in the medical domain. These tasks cover diagnosis, therapy planning, interacting with patients, identifying medical errors etc. Among these tasks, medical diagnosis has been one of the most popular research subjects in both medical informatics and computer science communities. [5]
- When guidelines are available, they provide a general framework to guide clinicians, but require consequent background knowledge to become operational, which is precisely the kind of information recorded in practice cases; cases complement guidelines very well and help to interpret them [4].
- Highly data intensive field in medicine, where it is advantageous to develop a system capable of reasoning from pre-existing cases from an electronic medical record, for instance, or from cases mined from the data. [4].

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 [4], [6], [7]. However, the medical applications offer a number of challenges for the CBR researchers and drive advances in research [8].

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.

II. 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 [9]. 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. 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) and t-distributed stochastic neighbor embedding (t-SNE).

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-paramatric density-based classfier.

III. 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. 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 prolongued 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. Consisting of 13 attributes which include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels coloured, thal 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*: 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.
- *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.

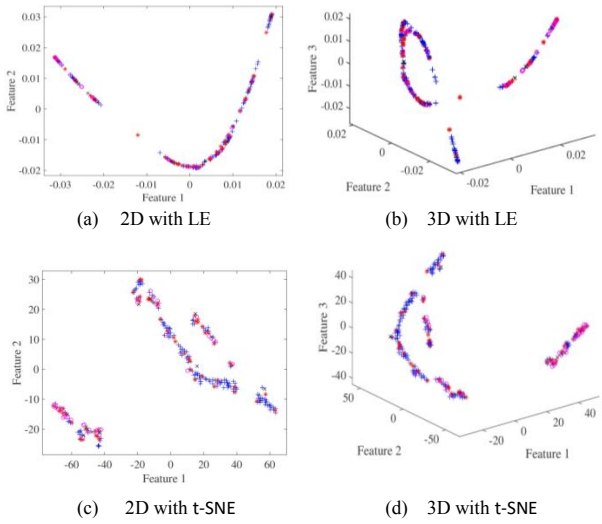
- *SVM*: This instance-based classification method takes advantage of the kernel trick to compute the most discriminative non-linear hyperplane between classes. Therefore, its performance heavily depends on the selection and tuning of the kernel type. For this work a Gaussian kernel is selected given its ability of generalization and its band-width parameter was fixed by the Silverman's rule [10].
- *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.

IV. 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 [11], where the classification accuracy with 7 attributes is of 70.36%.

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,



(a) 2D with LE (b) 3D with LE

(c) 2D with t-SNE (d) 3D with t-SNE

Fig. 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.

however, as for Cleveland dataset, dimensionality reduction does not substantially improves 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 [12] where they achieved an average accuracy of 0.9328.

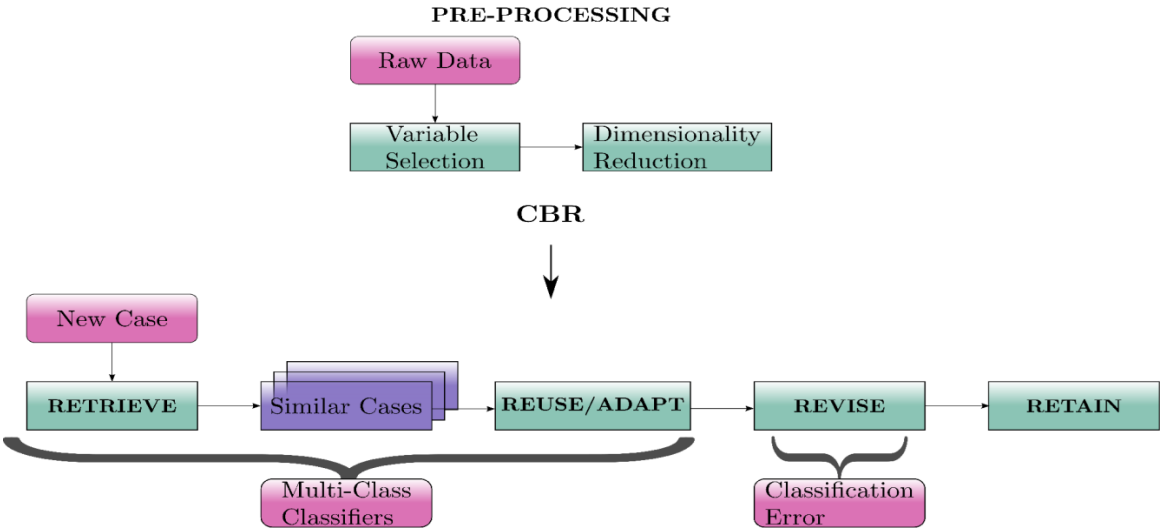
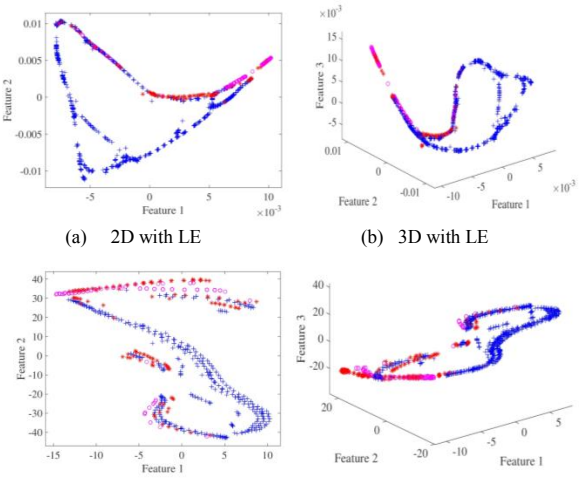


Fig. 1. Block diagram of proposed methodology. The aim of the comparative study is assessing the possibility of incorporating multi-class classifiers into CRB approaches design, as well as identifying the best classifier for this task.

TABLE I
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

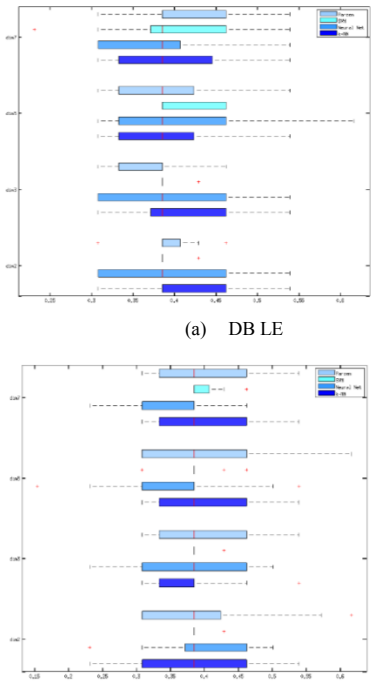
By performing a stability assessment, it could be seen from Figures 4,5 by the width of the error boxplots how SVM and *K*-NN classifiers achieves the best results for considered Cardiotocogram and Cleveland databases. Moreover, it should be noted how SVM classification results are the most stable of the considered classification techniques.



(a) 2D with LE (b) 3D with LE

(c) 2D with t-SNE (d) 3D with t-SNE

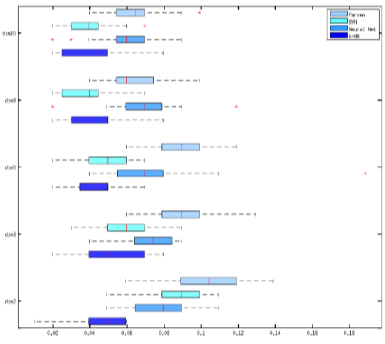
Fig. 3. Low-dimensional scatterplots for Cardiotocograms database. Figures (a), (c) show the first two features from database. Figures (b), (d) show the first three features from database.



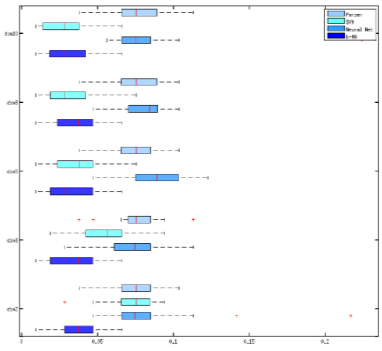
(a) DB LE

(b) DB t-SNE

Fig. 4. Classification error boxplots for considered classification techniques on Cleveland databases.



(a) DB t-SNE



(b) DB t-SNE

Fig. 5. Classification error boxplots for considered classification techniques on Cardiotocograms databases.

V. 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.

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 [4], [13].

As a future work, we will explore the possibility to design a case recovery stage for CBR able to deal with mult-class cases while providing users with class membership (probabilities to belong) estimates for a new case.

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

[1] L. Huan, X. Li, D, Hu, T, Hao, L. Wenyin and X, Chen. Adaptation Rule Learning for Case-Based Reasoning. " *Concurrency and Computation: Practice and Experience* ", 21(5), 673-689, 2009.

[2] J. Kolodner, Case-based Reasoning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[3] J. M. Juárez Herrero, "Una aproximación multimodal al diagnóstico temporal mediante razonamiento basado en casos y razonamiento basado en modelos. aplicaciones en medicina*," Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 11, pp. 77–80, 2007.

[4] I. Bichindaritz, "Case-based reasoning in the health sciences: What's next?" *Artificial Intelligence in Medicine*, vol. 36, no. 2, pp. 127–135, feb 2006.

[5] HT. Wang and AU Tansel. MedCase: A Template Medical Case Store for Case-Based Reasoning in Medical Decision Support*. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 962-967. ACM, 2013.

[6] L. Gierl and R. Schmidt, "CBR in medicine," in Case-Based Reasoning Technology, From Foundations to Applications. Springer-Verlag: New-York, 1998, pp. 273–298

[7] S. Montani, "Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support," Appl. Intell., pp. 275–285, 2007.

[8] S. Begum, M. Uddin, P. Funk, N. Xiong and M. Folke. 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), 41(4), 421-434. (2011).

[9] J. L. Kolodner, "Maintaining organization in a dynamic longterm memory," *Cognitive Science*, vol. 7, no. 4, pp. 243–280, 1983.

[10] S. J. Sheather *et al*., "Density estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, 2004.

[11] S. Bhatia, P. Prakash, and G. N. Pillai, "Svm based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features," 2008.

[12] Sundar. C, M. Chitradevi, and G. Geetharamani, "Article: Classification of cardiotocogram data using neural network based machine learning technique," *International Journal of Computer Applications*, vol. 47, no. 14, pp. 19–25, June 2012, full text available.

[13] M. Kwiatkowska and S. Atkins, "Case representation and retrieval in the diagnosis and treatment of obstructive sleep apnea: A semiofuzzy approach," 2004.