Using genetic algorithms to learn a fuzzy based pseudometric for k-NN classification

F. Martins[1], J. C. Becceneri[1], L. Dutra[1], D. Lu[2], and S. Sandri[1]

[1] Instituto Nacional de Pesquisas Espaciais
12201-970, São José dos Campos, SP, Brazil
E-mail: flavinha@dpi.inpe.br, becce@lac.inpe.br, luciano.dutra@dpi.inpe.br,
sandra.sandri@inpe.br
[2] Michigan State U. - Center for Global Change and Earth Observations, East Lansing, MI, USA
E-mail: ludengsh@msu.edu

Abstract. We address the derivation of pseudometric based on fuzzy relations for classification applications, by the use of genetic algorithms to learn the fuzzy relations. We present an experiment for the classification of land use in an area of the Brazilian Amazon region.

Keywords: k-NN classification, fuzzy partitions, genetic algorithms

1 Introduction

In a previous work [2], we proposed a a function called f^+, based on fuzzy relations, which are themselves derived from fuzzy partitions, for use in classification applications. This function is the complement in $[0, 1]$ of a particular kind of fuzzy relation, called an Order Compatible Fuzzy Relation (OCFR$_\preceq$), defined using a total order $(\Omega, \preceq)$ [10]. An OCFR$_\preceq$ itself is derived from a type of fuzzy partition (a collection of fuzzy sets), called Convex Fuzzy Partitions (CFP$_\preceq$). The creation of OCFR$_\preceq$ was motivated by the need to ease the burden of creating suitable relations for use the particular fuzzy case-based reasoning classification approach proposed in [8]. In [2], we proved that f^+ function is i) a pseudometric, when obtained from a specific type of CFP$_\preceq$, called 2-Ruspini, and, in particular, a ii) metric, when this CFP$_\preceq$ is moreover composed solely of triangular fuzzy sets. The same happens in the case of multidimensional domains, for function $f^+_{(\mu)}$ that aggregates the results obtained for f^+ in each domain, using the arithmetic means as aggregation operator μ.

Here we address the derivation of f^+ for k-NN classification applications [11], by the use of fuzzy genetic algorithms [1] to learn the fuzzy relations. We describe an application in the classification of land cover and use in an area of the Brazilian Amazon region.

2 Fuzzy relation based pseudometrics f^+ and $f^+_{(\mu)}$

Let $S : \Omega^2 \to [0, 1]$ be a fuzzy binary relation and $(\Omega, \preceq)$ be a total order. Formally, S is an Order Compatible Fuzzy Relation with Respect to a Total Order $(\Omega, \preceq)$ (*OCFR*$_\preceq$ or *OCFR*, for short), when it obeys the following properties [10]:

- $\forall x, y, z \in \Omega$, $S(x, x) = 1$ (*reflexivity*)
- $\forall x, y, z \in \Omega$, $S(x, y) = S(y, x)$ (*symmetry*)
- $\forall x, y, z \in \Omega$, if $x \preceq y \preceq z$, then $S(x, z) \leq \min(S(x, y), S(y, z))$ (*compatibility with total order* $(\Omega, \preceq)$, or $\preceq$*-compatibility* for short).

Let $(\Omega, \preceq)$ be a total order and let $\mathbf{A} = \{A_1, ..., A_t\}$ be fuzzy partition (a collection of fuzzy sets) in Ω; here A_i denotes a fuzzy set but also its associated membership function. Let the core and support of a fuzzy set A be defined as $core(A) = \{x \in \Omega \mid A(x) = 1\}$ and $supp(A) = \{x \in \Omega \mid A(x) > 0\}$), respectively [3]. Formally, $\mathbf{A}$ is a Convex Fuzzy Partition with Respect to a Total Order $(\Omega, \preceq)$ (*CFP*$_\preceq$ or *CFP*, for short), if it obeys the following properties [10]:

1. $\forall A_i \in \mathbf{A}, \exists x \in \Omega, A_i(x) = 1$ (*normalization*),
2. $\forall x, y, z \in \Omega, \forall A_i \in \mathbf{A}$, if $x \preceq y \preceq z$ then $A_i(y) \geq \min(A_i(x), A_i(z))$ (*convexity*),
3. $\forall x \in \Omega, \exists A_i \in \mathbf{A}, A_i(x) > 0$ (*domain-covering*),
4. $\forall A_i, A_j \in \mathbf{A}$, if $i \neq j$ then $core(A_i) \cap core(A_j) = \emptyset$ (*non-core-intersection*).

Let $\mathcal{A}_{(\Omega,\preceq)}$ denote the set of all CFPs that can be derived considering a total order $(\Omega, \preceq)$. CFP $\mathbf{A} \in \mathcal{A}_{(\Omega,\preceq)}$ is said to be a n*-CFP* if each element in Ω has non-null membership to at most n fuzzy sets in $\mathbf{A}$ ($n \geq 1$). In particular, a 2-CFP$_\preceq$ $\mathbf{A}$ is called a *2-Ruspini* partition, when it obeys additivity:

- $\forall x \in \Omega, \sum_i A_i(x) = 1$ (*additivity*)

In [10], the authors propose to generate OCFR$_\preceq$ $S^+ : \Omega^2 \to [0, 1]$ from a CFP$_\preceq$ $\mathbf{A}$ as

$$S^+(x, y) = \begin{cases} 0, \text{if } S^*(x, y) = 0 \\ S_L(x, y), \text{otherwise} \end{cases}$$

$$\forall x, y \in \Omega, S^*(x, y) = \sup_i \min(A_i(x), A_i(y))$$

$$\forall x, y \in \Omega, S_L(x, y) = \inf_i \; 1- \mid A_i(x) - A_i(y) \mid$$

Note that S_L is constructed based on the Lukasiewicz biresiduated operator [9].

In [2], the following function was proposed for tasks in which metrics and pseudometrics are employed[1]:

$$\forall x, y \in \Omega, f^+_{\mathbf{A}}(x, y) = 1 - S^+_{\mathbf{A}}(x, y).$$

This formula can be written directly as:

$$\forall x, y \in \Omega, f^+_{\mathbf{A}}(x, y) = \begin{cases} 1, \text{if } \forall i, \min(A_i(x), A_i(y)) = 0, \\ \sup_i \; \mid A_i(x) - A_i(y) \mid, \text{otherwise.} \end{cases}$$

[1] A metric satisfies non-negativity, symmetry and the triangle inequality and the identity of indiscernibles properties. Pseudometrics obey the same properties, except for the identity of indiscernibles, that is substituted by anti-reflexivity, a weaker property.

When no confusion is possible, we denote $f_{\mathbf{A}}^{+}$ as simply f^{+}.

Let $O = \Omega_1 \times ... \times \Omega_m$, where $\forall i, (\Omega_i, \preceq)$ is a total order. Let $\mathbf{A}_i$ be a 2-Ruspini CFP$_{\preceq}$ on Ω_i and f_i^{+} be derived from $\mathbf{A}_i$. Let $f_{(\mu)}^{+}\ :\ O \rightarrow [0,1]$ be the extension of function f^{+} to multidimensional domains, defined as

$$f_{(\mu)}^{+}(x,y) = \mu(f_1^{+}(x_1,y_1), ..., f_m^{+}(x_m,y_m)),$$

where $\mu : [0,1]^m \rightarrow [0,1]$ is the arithmetic mean, i.e., $\mu(a_1,...,a_m) = \frac{\sum_{1 \leq i \leq m} a_i}{m}$.

In [2], it is proved that $f_{\mathbf{A}}^{+}$ is a pseudometric, in general, and a distance when all fuzzy sets in $\mathbf{A}$ are triangular. Function $f_{(\mu)}^{+}$ trivially satisfies symmetry, anti-reflexitivity and non-negativity. The same result holds for $f_{(\mu)}^{+}$. In the same work, function $f_{(\mu)}^{+}$ was tested in a real-world application and yielded very good results when compared to both the Euclidean and Mahalanobis distances.

3 Learning $f_{(\mu)}^{+}$ using genetic algorithms for k-NN classification

We propose to use genetic algorithms to learn the fuzzy partitions necessary for function $f_{(\mu)}^{+}$, which is also our fitness function. Here we consider classification by k-NN but other methods could be used.

Let $X = \{x_1, ...x_m\}$ be a set of variables, each of which defined in domain $\Omega_i = [l_i, u_i], i \in \{1, m\}$. We encode each chromosome as a sequence of m genes, each of which related to a variable in X. The i-th gene is a sequence of parameters $< p_1, ..., p_s >$, representing points in domain Ω_i for a Ruspini partition. The sequence is such that $p_i \leq p_{i+1}, 1 \leq i \leq s-1$. In a trapezoidal partition, the first (respec. last) fuzzy term will have $[l_i, p_1]$ (respec. $[p_s, u_i]$) as core and $[l_i, p_2]$ (respec. $[p_{s-1}, u_i]$) as support. In a triangular partition, the first (respec. last) fuzzy term will have l_i (respec. u_i) as core and $[l_i, p_1]$ (respec. $[p_s, u_i]$) as support.

Crossover consists in choosing a cutting place in two selected chromosomes c_1 and c_2, and generating two new chromosomes c_{12} and c_{21}. Let chromosome c_i be described as $< p_{i,1}, ..., p_{i,s} >$ and let the cutting happen between the (k)-th and (k+1)-th genes. The crossover between any two chromosomes c_1 and c_2 would be generate two new chromosomes c_{12} and c_{21}, respectively described as $< p_{1,1}, ..., p_{1,k}, p_{2,k+1}, , ..., p_{2,s} >$ and $< p_{2,1}, ..., p_{2,k}, p_{1,k+1}, , ..., p_{1,s} >$

If one of the generated chromosomes does not satisfy the condition on the p_is, we reorganize the parameters. For example, let us suppose we have two chromosomes with 3 trapezoidal fuzzy sets Let c_1 and c_2 be described as $< 10, 20, 30, 40 >$ and $< 31, 32, 33, 34 >$, respectively, and that the cutting point is between p_2 and p_3. We obtain a valid chromosome, $c_{12} =< 10, 20, 33, 34 >$, and an invalid one, $c_{21} =< 31, 32, 30, 40 >$. We then rearrange the invalid chromosome as $c_{21} =< 30, 31, 32, 40 >$.

In this work we use n-fold cross-validation. First of all, a data set T is partitioned in n (approximately) equal parts (folds) T_i, such that $T = \cup_i T_i$. Then, for a given fold i, training is performed using the elements of all folds, except for those in i, and testing is performed the elements of fold i itself, making $Train_i = \bigcup_{T_j \in T, j \neq i} T_j$, and $Test_i = T_i$.

4 Experiments

In the following, we briefly describe an experiment that illustrates the use of function $f^{+}_{(\mu)}$ in a land use classification task in the Brazilian Amazon region. The area of interest covers approximately 411 km^2 and in the municipality of Belterra, state of Pará, in the Brazilian Amazon region, partially contained in the National Forest of Tapajós. An intense occupation process occurred in the region along the BR-163 highway (Cuiabá-Santarém), with opening of roads to establish small farms, after deforestation of primary forest areas [4]. As a result, there are mosaics of secondary vegetation in various stages, with pastures and cultivated areas embedded in a forest matrix [5].

In this application, 14 attributes have been considered, derived from either radar or optical satellite images, with 6 classes: forest, initial or intermediate regeneration, advanced regeneration or degraded forest, cultivated area, exposed soil, and pasture. The samples consist of 138 ground information based hand-made polygons. The attribute value for each polygon is the average of the values for the pixels composing it. The experiments have been done using 6 folds (5 for training and 1 for testing).

To obtain the lower (respec. upper) bound for a variable domain, we took the smallest (respec. largest) value from the elements in the fold, less (respec. plus) 20%. We have tested two types of partition for each variable, a triangular and a trapezoidal one, each of which with 3 fuzzy terms. In the triangular experiment, each partition is described by $< p_1 >$, where p_1 is the core of the middle triangular fuzzy term. In the trapezoidal experiment, each partition is described by $< p_1, p_2, p_3, p_4 >$, where $[p_2, p_3]$ is the core of the middle trapezoidal fuzzy term.

In our experiments, for each fold, the candidate population has 10 chromosomes. Each chromosome has 3 genes, each of which describing a partition corresponding to one of 3 variables used here. We have used an elitist genetic algorithm, keeping the best 6 elements and combining the 3 first elements to generate the new candidates that replace the worst 4 elements. We used a mutation rate of .2 and 400 generations.

We have used two kinds of population in the initial generation for each fold: "random" and "selected". In the selected first population for the fuzzy terms, the points are obtained from a fixed set of percentage vectors. Considering all domains to be normalized to [0,1], the selected population for the trapezoidal fuzzy sets corresponds to the set of 10 quadruples $< .20, .40, .60, .80 >$, $< .05, .28, .52, .76 >$, $< .23, .47, .71, .95 >$, $< .23, .47, .52, .76 >$, $< .23, .28, .52, .76 >$, $< .23, .47, .71, .76 >$, $< .4, .55, .7, .85 >$, $< .15, .55, .7, .85 >$, $< .15, .3, .7, .85 >$ and $< .15, .3, .45, .85 >$. The selected population for the triangular fuzzy sets is obtained by taking the arithmetic means between p_2 and p_3 from the trapezoidal fuzzy terms. It corresponds to $< .50 >$, $< .40 >$, $< .59 >$, $< .49 >$, $< .40 >$, $< .59 >$, $< .62 >$, $< .62 >$, $< .50 >$ and $< .37 >$.

Figure 4 brings the accuracy results for this application, considering k-NN with 1 to 6 neighbours, using the several versions of function $f^{+}_{(\mu)}$: trapezoid-based and triangle-based, considering selected and random initial populations (kNN_dFtz_s, kNN_dFtz_r, kNN_dFtg_s, kNN_dFtg_r). For comparison, the figure also brings the Euclidean distance (kNN_dE).

We see from the figures that all methods had high accuracy and that the best average results in the 6 folds were obtained with the use of f^{+}_{μ} for the triangular partitions. The best individual results, considering all folds, were the same methods for 1, 2 and 3

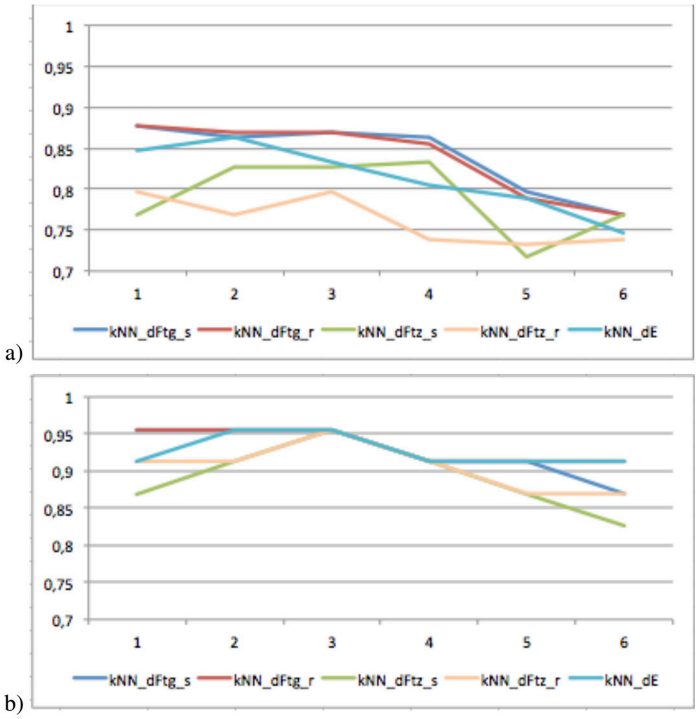

Fig. 1. Classification accuracy results for: a) k-NN average and b) k-NN maximum.

neighbours and the Euclidean distance for 2 and 3 neighbours. In particular, f_μ^+ for the triangular partitions with the initial population obtained at random yielded the same results for the maximum as the Euclidean distance, except for 1 neighbour, when f_μ^+ fares better. All methods fare better with a small number of neighbours. In particular, the best results for the triangular partitions, considering both the average and the maximum, is obtained already with a single neighbour. The worst results have been obtained with the trapezoidal partitions, for both types of initial populations.

5 Conclusions

In this work, we have proposed to use of genetic algorithms to learn fuzzy relations, that are parameters for a pseudometric $f_{(\mu)}^+$. We describe a classification application of land cover and use in an area of the Brazilian Amazon region, using k-NN. The results have shown that the triangular partitions produced the best results.

Future work includes experimenting with other data sets. We also intend to verify alternatives to reduce the computational cost, without a decrease in accuracy or adequately reducing the training data Another alternative consists in learning the partition for each variable separately; in order to calculate accuracy the distance relative to the other variables would be fixed (e.g. Euclidean) and aggregated with the distance obtained from the partition.

This work is a first step towards using $f^{+}_{(\mu)}$ in [7], an extension to k-NN for image classification, in which there is the possibility of using multiple spaces, that can be originated from different data sources, having different ranges of values, as well as the geographical space itself, allowing the use of topological associations.

Acknowledgements The authors acknowledge financial support from Brazilian research funding agencies CAPES, CNPq and FAPESP.

References

1. F. Herrera, Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions, International Journal of Computational Intelligence Research, Vol.1, No.1, pp. 59-67 (2005)
2. Sandri, S., Martins-Bedê, F.T., Dutra, L.,: Using a Fuzzy Based Pseudometric in Classification. Proc. 14th Int. Conf. on Info. Proc. and Management of Uncertainty in Knowledge-Based Systems, IPMU 2014, Montpellier-Fr (2014)
3. Dubois, D., Prade, H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
4. Brazilian Institute of Environment and Renewable Natural Resources (IBAMA): Floresta Nacional do Tapajós Plano de Manejo. Vol I. (in Portuguese) Available at: <http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/flona_tapajoss.pdf>. Date accessed: 30 Dec. 2013 (2009)
5. Escada, M.I.S., Amaral, S., Rennó, C.D., Pinheiro, T.F.: Levantamento do uso e cobertura da terra e da rede de infraestrutura no distrito florestal da BR- 163. Repport INPE-15739-RPQ/824, INPE, S.J.Campos, Brazil. Available at: <http://urlib.net/sid.inpe.br/mtc-m18@80/2009/04.24.14.45>. Date accessed: 30 Dec. 2013 (2009)
6. Korsrilabutr, T., Kijsirikul, B.: Pseudometrics for Nearest Neighbor Classification of Time Series Data. Engineering Journal, Thailand, 13, May. 2009. Available at: <http://engj.org/index.php/ej/article/view/46>. Date accessed: 30 Dec. 2013 (2009)
7. Martins-Bedê, F.T.: Souza Reis, M., Pantaleão, E., Dutra, L., Sandri, S. An application of multiple space nearest neighbor classifier in land cover classification Proc. IGARSS'14, pp 1713–1716 (2014)
8. Mendonça, J.H., Sandri, S., Martins-Bedê, F.T., Guimarães, R., Carvalho, O.: Training strategies for a fuzzy CBR cluster-based approach, Mathware & Soft Computing, v. 20, pp 42–49 (2013)
9. Recasens, J.: Indistinguishability operators, modelling fuzzy equalities and fuzzy equivalence relations. Series: Studies in Fuzziness and Soft Computing, vol 260, Springer Verlag (2011)
10. Sandri, S., Martins-Bedê, F.T.: A method for deriving order compatible fuzzy relations from convex fuzzy partitions. Fuzzy Sets and Systems, pp 91–103 (2014)
11. Theodoridis, S. and Koutroumbas, K.: Pattern recognition, Academic Press, 3rd edition (2006)