GFS-Based Analysis of Vague Databases in High Performance Athletics

Ana Palacios[1], Inés Couso[2], and Luciano Sánchez[1]

[1] Universidad de Oviedo, Departamento de Informática, 33071 Gijón, Asturias, Spain
apalaciosgimenez@gmail.com, luciano@uniovi.es,
[2] Univ. de Oviedo, D. de Estadística e I.O. y D.M., 33071 Gijón, Asturias, Spain
couso@uniovi.es

Abstract. To configure a proficient athletics team, coaches combine their expertise with the analysis of data collected during training sessions and competitions. This data and knowledge are vague, thus fuzzy logic is appropriate for designing a decision model. In this paper we will use a Genetic Fuzzy System for designing a model that can help the trainer to assess the performance of a given athlete in the future, given a combination of historical data and expert knowledge. This decision model has interest as a real-world application of GFSs, but it also involves novel kinds of data, whose study is a current trend in machine learning. Examples of such data include subjective perceptions of mistakes of the athletes, the reconciliation of different measurements taken by different observers, and interval-valued training data. We will use a possibilistic representation of these categories of information, in combination with an extension principle-based reasoning method, and finally show that the quality of a GFS which is based in these last principles improves the results of the original formulation of the same algorithm.

1 Introduction

The most important decision of a professional atlethics coach is that of selecting the team that will take part in a competition. There are many constraints involved in that decision; let us remark two: only a limited number of athletes can be called, and each one of these athletes must participate in a number of different races. Since the objective in the selection is obtaining the highest score for the whole team, the individual capabilities must be balanced. If the marks that each athlete will obtain at each race could be known in advance, then the composition of the best team would be immediate [2]. However, to the best of our knowledge, the design of an athletics team has not been automatized yet, and there are not previous works about intelligent models of the performance of athletes.

According to our own research, most coaches believe that an accurate prediction of the performance of an athlete at a future event is not possible, and they rely instead in a simpler, threshold-based mechanism. They establish a baseline mark, and decide whether an athlete will be able to improve that mark or not [9]. The selection of those baseline marks is not a trivial subject. Time ago, it was common that a trainer set an unique mark for the whole team, that depended on the results of the rival teams. By

E. Corchado and H. Yin (Eds.): IDEAL 2009, LNCS 5788, pp. 602–609, 2009.

contrast, the current trend is to select a different mark for each athlete [9], let it be the best personal mark of the athlete, the regional record, or other value that serves the coach to decide whether the athlete is needed. Once that baseline mark is settled, predicting whether this mark will be reached or not is a complex decision. A coach uses his expertise, his personal knowledge of the athletes and also observes a number of indicators, which are either numbers, words, interval (or fuzzy) ranges of values, or compound measures.

In this paper, we propose a method for discovering the linguistic rules that model the expertise of a coach, by mining a database that contain the past performance of the athletes, and the values of the aforementioned indicators. We will use a Genetic Fuzzy System to do this task. Furthermore, we have used a nonstandard model that accounts for the imprecision in the indicators, which is based on the theory of possibility. Our representation of the data is explained in Section 2. This representation requires some changes in the inference procedure, that will be reviewed in Section 3, and also some changes in the genetic algorithm, summarized in Section 4. In Section 5 we explain the structure of the decision model, and review the indicators on which it depends. Lastly, in Section 6, we setup two Genetic Fuzzy Systems that only differ on the representation of the data and the inference mechanism, and show that the changes proposed in this paper account for a better prediction capability.

2 Possibilistic Representation of Vague Information

Possibilistic representations of vague information are commonly used in fuzzy statistics, but they are not so common in fuzzy logic-based models. We are interested in those cases where we cannot accurately observe all the properties of the object, but we will be given a nested family of sets, each one of them containing the true value with certain probability. This apparently complex specification matches well with a large amount of practical situations: for instance, it includes datasets with missing values (one interval that spans the whole range of the variable), left and right censored data (the value is higher or lower than a cutoff value, or it is between a couple of bounds), compound data (each item comprises a disperse list of values), mixes of punctual and set-valued measurements (as produced by certain sensors, for instance GPS receivers) etc. All these cases share a certain degree of ignorance about the actual value of a variable, thus we will refer to them with the generic term “low quality data”.

Recent works in fuzzy statistics suggest using a fuzzy representation when the data is known through a family of confidence intervals [1]. This representation assumes that a fuzzy set can be interpreted as a possibility distribution (which, in turn, is a family of probability distributions) and each α-cut of a fuzzy feature is a random set that contains the unknown crisp value of the feature with probability $1 - \alpha$ (see [12,14] and Figure 1). The adoption of this representation is not, however, compatible with other interpretations of a fuzzy set, that must be modified in accordance, as we will discuss in the next section.

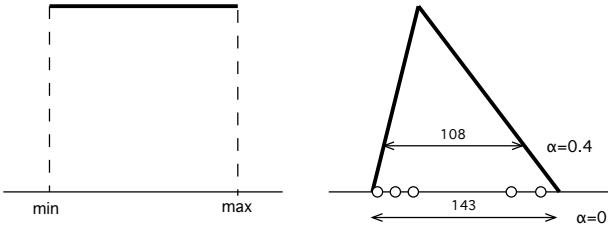

Fig. 1. Fuzzy representation of low quality data. Left: A missing value is codified with an interval that spans the whole range of the variable, or $P([\min, \max]) \leq 1$. Right: A compound value (in this example, five different measurements of the variable) can be described by a fuzzy membership, that can also be understood as an upper probability. Each α-cut contains the true value of the variable with probability at least $1 - \alpha$.

3 An Extension Principle-Based Reasoning Method

There is one important caveat about the combination of the possibilistic representation introduced in the preceding paragraphs and a fuzzy logic-based model or classifier: not all the reasoning methods preserve the possibilistic meaning of the data. That is to say, for most definitions of fuzzy inference, it may happen that, given an input that has a possibilistic meaning, we come out with an output that has not that kind of interpretation. In order to obtain meaningful results, in this section we adapt a reasoning method, that was proposed in [14] for fuzzy models, to the classification case.

Let X be the input space, let N_c be the number of classes, thus $K = \{1, \ldots, N_c\}$ is the output space, and let $\{\widetilde{A}_i \rightarrow C_i\}_{i=1,\ldots,M}$ be a set of M fuzzy rules. Given an input $x \in X$, the most common reasoning method for computing the output of a FRBS takes two stages:

1. An intermediate fuzzy set is composed:

$$\widetilde{\text{out}}(x)(k) = \max_{i=1,\ldots,M} \min\{\widetilde{A}_i(x), \delta^k_{C_i}\}. \tag{1}$$

2. This intermediate fuzzy set is transformed in a crisp value $\text{defuz}(\widetilde{\text{out}}(x)) \in K$ by means of a suitable defuzzification operator.

Given a set-valued input $A \subseteq X$ (that, in our context, means "all we know about the input is that it is in the set A") we propose to operate as follows:

1. We determine a family of intermediate fuzzy sets in the universe $\mathcal{F}(K)$, $\widetilde{\text{out}}(A) \in \wp(\mathcal{F}(K))$, defined as

$$\widetilde{\text{out}}(A) = \{\widetilde{\text{out}}(x) \text{ s. t. } x \in A\} \tag{2}$$

2. An element of $\wp(K)$ (that is to say, a set of crisp outputs $\text{defuz}(\widetilde{\text{out}}(A)) \in \wp(K)$) is obtained, according to the following definition:

$$\text{defuz}(\widetilde{\text{out}}(A)) = \{\text{defuz}(\widetilde{\text{out}}(x)) \text{ s. t. } x \in A\}. \tag{3}$$

Lastly, given a fuzzy input $\widetilde{A} \in \mathcal{F}(X)$, we will assign it, according to the Extension Principle (which is compatible with the possibilistic interpretation of fuzzy sets) a fuzzy set computed as follows:

1. We determine an intermediate fuzzy set on the universe $\mathcal{F}(K)$, $\widetilde{\text{out}}(\widetilde{A}) \in \mathcal{F}(\mathcal{F}(K))$, defined as

$$\widetilde{\text{out}}(\widetilde{A})(\widetilde{B}) = \sup\{\widetilde{A}(x) \text{ s. t. } \widetilde{\text{out}}(x) = \widetilde{B}\}, \quad \forall \widetilde{B} \in \mathcal{F}(K) \tag{4}$$

2. An element of $\mathcal{F}(K)$ (that is to say, a fuzzy output) $\text{defuz}(\widetilde{\text{out}}(\widetilde{A})) \in \mathcal{F}(Y)$ is obtained as follows:

$$\text{defuz}(\widetilde{\text{out}}(\widetilde{A}))(k) = \sup\{\widetilde{A}(x) \text{ s. t. } \text{defuz}(\widetilde{\text{out}}(x)) = k\}, \quad \forall k \in K. \tag{5}$$

Observe that the fuzzy set $\text{defuz}(\widetilde{\text{out}}(\widetilde{A}))$ is associated to the nested family of sets $\{\text{defuz}(\widetilde{\text{out}}(\widetilde{A}_\alpha))\}_{\alpha \in [0,1]}$, and that explains the possibilistic interpretation of this procedure.

4 Obtaining and Validating a Fuzzy Rule-Based Classifier from Low Quality Data

Roughly speaking, estimating a classifier from data requires a numerical technique that finds the minimum of the classification error with respect to the free parameters of the classifying system. In our case, this function is fuzzy-valued. But there are not many techniques for optimizing interval-valued [3] or fuzzy valued functions. In the genetic algorithms field, the solutions are related to precedence operators between imprecise values [5,6,15]. We have previous works where we have jointly optimized a mix of crisp and fuzzy objectives with genetic algorithms [12]. We have also proposed a number of different algorithms for learning regression models from low quality data and the fuzzy representation mentioned before [11,13,14]. Recently, we have also proposed two different algorithms for classification problems [7] and also for risk-based classification [8]. In this paper we will use the algorithm defined in [7], which we will not detail again, because of space limitations. This algorithm is an extension of the Genetic Cooperative-Competitive Learning introduced in [4] to a possibilistic representation, using a particular case of the reasoning mechanism introduced in the preceding section, where

$$\text{defuz}(\widetilde{\text{out}}(x)) = \arg\max_k \{\widetilde{\text{out}}(x)(k)\}. \tag{6}$$

This GFS is a good choice for evaluating the impact of the possibilistic representation in the problem of High Performance Athletics. It is an straightforward generalization of the crisp algorithm in [4], thus all the improvements in the numerical results with respect to that algorithm (see Section 6) can be attributed to the new representation and reasoning mechanisms.

5 Structure of the Proposed Fuzzy Rule-Based Decision Model

The score of an athletics team is the sum of the individual scores of the athletes in the different events. As we have mentioned before, it is the coach's responsibility to balance the capabilities of the different athletes in order to maximize the score with a team according to the regulations [10].

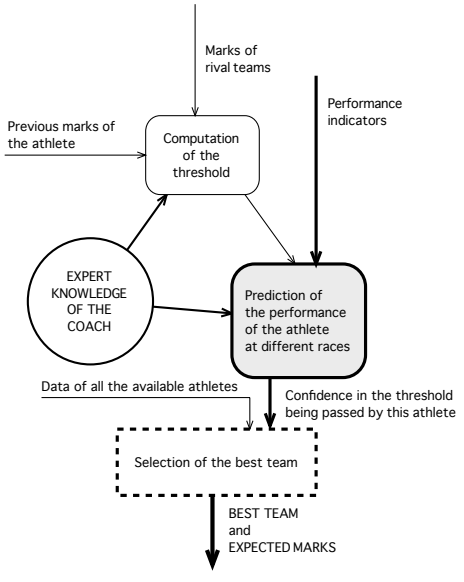

Fig. 2. Structure of the model

In Figure 2 we have summarized the steps of this process. The algorithm that is used to choose the team is fed with the expected marks ("thresholds") of the athletes at each event, and also with confidence degrees in the achievement of these marks. These thresholds are determined by the trainer according to the past performance of the athlete, a set of indicators that are described in this section and, optionally, the marks of rival teams. The objective is to choose the subset of athletes that will get a given score, with the highest confidence [2]. In this paper we are interested in the prediction of the threshold (grey box in Figure 2).

The different events where the team will collaborate are divided into sprints, middle distance, long distance running, hurdling, relays, road, jumps and throwing. Each event has, in turn, different categories. For instance, there are 100 metre hurdles and 400 metre hurdles; both are speed races. In the next section we will review the indicators used in two of these events: long jump and 100 metres.

5.1 Long Jump

There are four indicators in long jump that are used to predict whether an athlete will pass a given threshold [16]: the ratio between the weight and the height, the maximum speed in the 40 metre race and the tests of central (abdominal) muscles and lower extremities. The first two indicators are determined by the coach, who was allowed to use numbers, intervals or linguistic values (fuzzy intervals) at his convenience. The two last tests are repeated three times, and produce numbers. The abdominal muscle test consists in counting how many flexion movements the athlete can repeat in a minute. Lastly, the lower extremities test measures how much the athlete can stretch.

5.2 100 Metre Sprint

There are also four indicators in this event: the ratio between weight and height, the reaction time, the starting or 20 metre speed, and the maximum or 40 metre speed.

We have collected two different databases for this problem. In the first database, three different people measure the actual reaction time, starting and maximum speed of the athletes. These three measurements are joined to form an imprecise value. On the contrary, in the second database the trainer has graded each speed and time with a mark between 0 and 10. He was allowed to express his grades with numbers, intervals or linguistic values. This second database has a high subjective component; it serves to assess the expert knowledge of the trainer about the athletes, by comparing this results with the actual measurements.

6 Numerical Analysis of the Algorithm

In this section we have compared the results of a GFS that uses crisp datasets to the same GFS, but extended to use possibilistic data, with the representation and inference function we have mentioned in the preceding sections. All our studies have been carried with athletes of the Oviedo University that participate in the Spanish Women's Athletic Club Championship. We have collected three datasets, whose description is as follows:

6.1 Description of the Datasets

1. Dataset "Long-4": This dataset is used to predict whether an athlete will improve certain threshold in the long jump, given the indicators mentioned before. We have measured 25 athletes, thus the set has 25 instances, 4 features, 2 classes, no missing values. All the features, and also the output variable, are interval-valued.
2. Dataset "100ml-4-P": Used for predicting whether a threshold in the 100 metres sprint race is being achieved. Actual measurements are taken by three observers, and are combined into the smallest interval that contains them. 25 instances, 4 features, 2 classes, no missing data. All input and output variables are intervals.
3. Dataset "100ml-4-I": Same dataset as "100ml-4-P", but the measurements have been replaced by the subjective grade the trainer has assigned to each indicator (i.e. "reaction time is low" instead of "reaction time is 0.1 seg").

6.2 Compared Results

We have compared the performance of the generalized algorithm to that of the original crisp algorithm. To that end, we have built a crisp dataset by removing the uncertainty in the imprecise dataset: each imprecise measurement was replaced by the mid-point of the corresponding interval, and those examples with imprecision in the independent variable were replicated for the different options.

For instance, a point $(X = [1, 3], C = \{A, B\})$ is converted into two points $(x = 2, c = A)$, $(x = 2, c = B)$. In this case, each one of the instances is assigned a weight equal to the inverse of the number of duplicates, so that the contribution of the instance to the global error does not change. In other words, to defuzzify the vague data we have assumed a prior knowledge about the uncertainty of the measurements: the uniform probability distribution.

We have used a 10cv design for all datasets. The boxplots with all the results are shown in Figure 3. Observe that the boxplots of the imprecise experiments are not

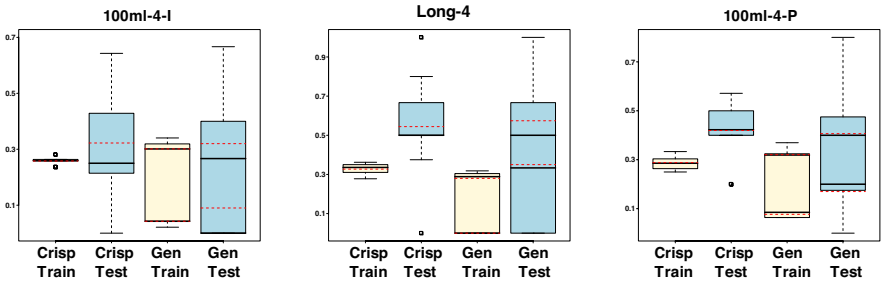

Fig. 3. Boxplots illustrating the dispersion of the 10 repetitions of crisp and generalized GFS. From left to right,: Problems "100ml-4-I", "Long-4", and "100ml-4-P".

Table 1. Means of 10 repetitions of the generalized GFS for the imprecise datasets "Long-4", "100ml-4-P" and "100ml-4-I"

	Crisp		Low Quality	
Dataset	Train	Test	Train	Test
Long-4 (5 labels)	0.327	0.544	[0.0,0.279]	[0.349,0.616]
100ml-4-P (5 labels)	0.288	0.419	[0.076,0.320]	[0.17,0.406]
100ml-4-I (5 labels)	0.259	0.384	[0.089,0.346]	[0.189,0.476]

standard. We propose using a box showing the 75% percentile of the maximum and 25% percentile of the minimum fitness (thus the box displays at least the 50% of data) and also drawing two marks inside the box to mark the interval-valued median. We have also included the mean value with a couple of red lines. The numerical values of the classification error have also been included in Table 1.

The results are promising in all the experiments. We expected that the extra freedom that the coach has when he is allowed to use ranges of values and linguistic terms instead of numbers allowed us to capture better his expertise, and the results seem to confirm this intuition (first boxplot in Figure 3, dataset 100ml-4-I). The other datasets show also a remarkable improvement, having into account that the algorithm used is an overly simple extension of a GFS to the use of low quality data. We expect to improve these results with future learning algorithms that are designed to take advantage of the new possibilistic representation.

7 Concluding Remarks

The work explained in this paper has a preliminary nature. The field work is finished, and data from 25 athletes has been captured, along with the subjective grading of all of them by the trainer. We have modeled the expertise of the coach with fuzzy rules, and shown that a possibilistic representation, combined with the extended inference, improves the quality of the modeling. However, there is still much room for improvement, as the dificulty of the problem is very high. The percentage of wrong classifications found by the GFS is still too high for this model being useful in practice.

In future works we will apply different machine learning algorithms to these datasets and make a compared analysis, and include new, more efficient GFSs that are based in the fuzzy-valued genetic algorithm described in [14].

Acknowledgements

This work was supported by Spanish Ministry of Science and Innovation, under grant TIN2008-06681-C06-04, and by Principado de Asturias, PCTI 2006-2009.

References

1. Couso, I., Sánchez, L.: Higher order models for fuzzy random variables. Fuzzy Sets and Systems 159, 237–258 (2008)
2. Chen, S.: Analysis of maximum total return in the continuous knapsack problem with fuzzy object weights. Applied Mathematical Modelling 33(7), 2927–2933 (2009)
3. Floudas, C.A., Pardalos, P.M.: Encyclopedia of Optimization. Springer, Heidelberg (2009)
4. Ishibuchi, H., Nakashima, T., Murata, T.: A fuzzy classifier system that generates fuzzy if-then rules for pattern classification problems. In: Proc. of 2nd IEEE International Conference on Evolutionary Computation, pp. 759–764 (1995)
5. Koeppen, M., Franke, K., Nickolay, B.: Fuzzy-Pareto-Dominance driven multi-objective genetic algorithm. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 450–453. Springer, Heidelberg (2003)
6. Limbourg, P.: Multi-objective optimization of problems with epistemic uncertainty. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 413–427. Springer, Heidelberg (2005)
7. Palacios, A., Sánchez, L., Couso, I.: A baseline genetic fuzzy classifier based on low quality data. In: EUSFLAT 2009 (accepted, 2009)
8. Palacios, A., Sánchez, L., Couso, I.: A minimum-risk genetic fuzzy classifier based on low quality data. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS (LNAI), vol. 5572, pp. 654–661. Springer, Heidelberg (2009)
9. Palacios Martín, J.L.: Comunicación personal (2009)
10. Escolar, A.G. (ed.): Reglamento Internacional de Atletismo (1995)
11. Sánchez, L., Otero, J., Villar, J.R.: Boosting of fuzzy models for high-dimensional imprecise datasets. In: Proc. IPMU 2006, Paris, France, pp. 1965–1973 (2006)
12. Sánchez, L., Couso, I., Casillas, J.: Modelling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria. In: Proc. 2007 IEEE Symp. on Comp. Int. in Multicriteria Decision Making, Honolulu, USA, pp. 30–37 (2007)
13. Sánchez, L., Otero, J., Couso, I.: Obtaining linguistic fuzzy rule-based regression models from imprecise data with multiobjective genetic algorithms. Soft Computing 13(5), 467–479 (2008)
14. Sánchez, L., Couso, I., Casillas, J.: Genetic learning of fuzzy rules based on low quality data. Fuzzy Sets and Systems, doi:10.1016/j.fss.2009.03.004 (in press)
15. Teich, J.: Pareto-front exploration with uncertain objectives. In: Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., Corne, D.W. (eds.) EMO 2001. LNCS, vol. 1993, pp. 314–328. Springer, Heidelberg (2001)
16. Vinuessa, M., Coll, J.: Tratado de atletismo. Servicio Geográfico del Ejercito Español (1984)