A feature selection algorithm with Fuzzy information

*Antonio Arauzo Azofra and Jose Manuel Benítez Sanchez and Juan Luis Castro Peña
Department of Computer Science and Artificial Intelligence (DECSAI)
University of Granada (Spain)
email: arauzo (a) decsai.ugr.es

Abstract

This paper deals with the problem of feature selection. Almuallim and Dieterich [1] developed the FOCUS algorithm which performs optimal feature selection on boolean domains. In a previous paper an extension of FOCUS is developed to deal with discrete and continuous features. In this paper we present an extension to work with fuzzy features, which is verified on two well known problem with quite good results.

1 Introduction

Feature selection help us to focus the attention of an induction algorithm in those features that are the best to predict a target concept. Although one might think that the more information available to an induction algorithm the better it works, this has revealed to be false for the following two main reasons. First, a large number of features in the input of induction algorithms may turn them very inefficient as memory and time consumers. And second, irrelevant data may confuse algorithms making them to reach false conclusions.

In feature selection, we are interested in finding the minimal set of features which allows us to induce the target concept. John, Kohavi and Pfleger[5] classify the features in three relevance classes: irrelevant, weakly relevant and strongly relevant. The FOCUS algorithm[1] is successful identifying the set with all strongly relevant and the minimal number of weakly relevant features to the target concept. As result of this, FOCUS is an ideal algorithm to use when a minimal set of features is required and noise free samples are available.

FOCUS always finds the optimal set through a complete search on the feature subset space in quasi-polynomial time. In [3] very interesting empirical results of the FOCUS algorithm are presented. It displays good performance even on some datasets with noise.

However FOCUS and FOCUS-2[1] (the optimized version of FOCUS) are limited to boolean domains, while many real problems have discrete and continuous attributes. In order to avoid this drawback C-FOCUS[2] was developed as a FOCUS-2 extension to deal with nominal, discrete and continuous features.

In this paper we consider how can C-FOCUS be used with fuzzy information, developing the extension F-FOCUS. There are three main reason why we think this algorithm will be useful. The first one is that although F-FOCUS is not a wrapper algorithm and it works independently from any classifier that could be used later, it has some of the advantages of embebed feature selection algorithms when used with fuzzy classification systems, given that it is using the same fuzzy feature definition. The second one is that in problems defined with continuous features we can introduce expert knowledge selecting the different fuzzy sets for every feature. And finally it allows, by reducing the number of features, to solve some problems with fuzzy rules that would not be solv-

*This research has been supported by project CICYT-TIC2000-1362-C02-01

able because of the exponential efficiency of fuzzy rule systems.

Another possible application is the use of F-FOCUS feature selection to help the classifier by introducing some derived features, for example products of fractions of features, and selecting those that are relevant to the problem. This application is shown in our second experiment.

In section 2, we describe FOCUS algorithm and its extension F-FOCUS with support of fuzzy features. In section 3, results of F-FOCUS application to two well known problems are shown. Finally, some conclusions and final remarks are collected in section 4.

2 Description of the Algorithm

The main idea of the original FOCUS algorithm is to identify all pairs of examples with a different boolean result. Each of these pairs is called a conflict, and FOCUS goal is to select the minimal set of features that solves all conflicts. A feature is considered to solve a conflict when its value is different between both examples. That is when the feature allow us to distinguish between the two examples.

It is clear when two values are different in a boolean or discrete domain, so it is clear when a conflict is solved by a boolean or discrete variable. But we need to define when two continuous values will be considered different.

To consider two values as distinct C-FOCUS utilizes the absolute difference between the two values in the following simple way. All values in samples of a given feature are normalized to $[0, 1]$. If the difference is greater than a given threshold U the two values will be considered distinct.

Both approaches for discrete and continuous values will continue to be used as in C-FOCUS for their respective feature types, and we extend here C-FOCUS to deal with fuzzy features.

To include fuzzy features we need to define when two fuzzy values are distinct. We have used a variation of the measure of separability described in [6] and considered that two values are distinct when this separability measure exceeds the given threshold U. The measure is described in the following equation.

$$d(x, y) = \sum_{L,L' \in FuzzyLabels} L(x)L'(y)D(L, L')$$

where $D(L, L')$ is the measure of separability of two fuzzy sets defined in [6]

FOCUS searches through the space of feature subsets to find the one with a minimal number of features that solves all conflicts.

This search can be done trying sequentially with all sets of $1, 2, 3, \ldots N$ variables until one set that solves all conflicts is found. But if one conflict is solved only by a feature X_i, we know that X_i should belong to the set of features selected. We design our algorithm as an extension of FOCUS-2 [1] which uses the aforementioned heuristic.

Algorithm FOCUS-2(*Sample*)

1. If all the examples in *Sample* have the same class, return $\emptyset$.
2. Let G be the set of all conflicts generated from *Sample*.
3. $Queue = \{M_{\emptyset,\emptyset}\}$.
4. Repeat
   - 4.1 $M_{A,B}$ = Pop the first element in *Queue*.
   - 4.2 $OUT = B$.
   - 4.3 Let a be the conflict in G not covered by any of the features in A, such that $|Z_a - B|$ is minimized, where Z_a is the set of features covering a.
   - 4.4 For each $x \in Z_a - B$
     - 4.4.1 If $Sufficient(A \cup \{x\}, Sample)$, return $A \cup \{x\}$.
     - 4.4.2 Insert $M_{A\cup\{x\},OUT}$ at the tail of *Queue*.
     - 4.4.3 $OUT = OUT \cup \{x\}$.

end.

$M_{A,B}$ denotes the space of all feature subsets that include all of the features in the set A and none of the features in the set B.

As the sufficiency test of step 4.4.1, $Sufficient(Features, Sample)$, we have used a

Table 1: Pima results

Test	F-FOCUS		All
0	"Times pregnant" "Plasma glucose" "Age"	68.18	77.92
1	"Plasma glucose" "Times pregnant" "Body mass"	79.87	79.87
2	"Plasma glucose" "Body mass" "Age"	72.73	78.57
3	"Plasma glucose" "Times pregnant" "Age"	73.20	81.05
4	"Plasma glucose" "Body mass" "Age"	82.35	75.82
5	"Plasma glucose" "DiastolicBP" "Age"	71.43	75.32
6	"Plasma glucose" "Body mass" "Age"	79.22	75.97
7	"Plasma glucose" "Body mass" "Age"	79.87	77.92
8	"Plasma glucose" "Times pregnant" "Body mass"	78.43	82.35
9	"Plasma glucose" "Body mass" "Age"	74.51	78.43
Mean		75.97	78.32

simple search through *Sample* of two examples, with values not considered different in selected *Features*, that belong to a different class. If there are no such two examples the *Features* set is sufficient, not being sufficient otherwise.

The difference between the two extensions(C-FOCUS, F-FOCUS) and FOCUS can be summarized in the way they handle values in features. This can easily be shown with the following example. Considering the feature cost: FOCUS will differentiate between only two values affordable and unaffordable. C-FOCUS will see the cost as a real number and will consider two values different when they are at more than a fixed distance. And finally F-FOCUS will consider the fuzzy values as cheap, medium, and expensive.

3 Empirical Study

As the values from the features of the datasets used are continuous and no fuzzy information was available, we have defined fuzzy sets for each feature: 5 equal sized trapezoidal fuzzy sets covering the domain of example values and 2 fuzzy sets to cover the rest of real domain.

3.1 Pima Indians Diabetes Database

This dataset belongs to the collection of UCI repository [4].

We have used cross-validation 80%-20% twice to generate ten training-test pairs. The results are shown in table 1. It displays the classification percentage obtained with one hidden layer perceptron using the selected features (second column) and using every feature (third column).

F-FOCUS has chosen Plasma glucose in all test and a combination of two features from Body mass, Age and Times pregnant. Using just these three features the classification percentage is just a bit lower than that with all the features.

3.2 Iris Plant Database

This very well known dataset has been widely used as a benchmark set. We have taken its 4 features and the product between each pair of these features in order to chose the most useful features among the original features or derived products. This introduces another way of using feature selection to help the classifier providing it more useful variables.

We have used cross-validation 80%-20% twice to generate ten training-test pairs. The results are shown in table 2. The classification percentage obtained running a one hidden layer perceptron with the selected features contrasted with those obtained with the same perceptron with all the features.

F-FOCUS has chosen petal or sepal length and one product in all cases. As can be checked, the results obtained employing just two features are slightly better that those reached using all of the features.

As it can be seen by the results obtained after

Table 2: Pima results

Test	F-FOCUS		All
0	"petalL" "petalLxpetalW"	96.67	86.67
1	"sepalL" "petalLxpetalW"	100.00	96.67
2	"sepalL" "petalLxpetalW"	93.33	90.00
3	"sepalL" "petalLxpetalW"	96.67	96.67
4	"sepalL" "petalLxpetalW"	96.67	100.00
5	"petalL" "sepalWxpetalW"	96.67	90.00
6	"petalL" "sepalWxpetalW"	96.67	100.00
7	"petalL" "petalLxpetalW"	96.67	93.33
8	"sepalL" "petalLxpetalW"	96.67	90.00
9	"petalL" "sepalWxpetalW"	90.00	96.67
Mean		96.00	94.00

feature selection with just two features are a bit better than those using the classifier directly.

4 Summary and Conclusions

We have developed the F-FOCUS algorithm as an extension of the C-FOCUS [2] algorithm to feature with fuzzy information. In this way, it can be applied to a wider set of problems.

This algorithm is recommended in classification problems in which we have noise free samples and the main goal is to reduce the number of features. We have tested the F-FOCUS algorithm on problems of both classes and obtained good results with reduced feature sets.

The first benefit of F-FOCUS is that it can help to apply fuzzy rule based systems to solve large dimension problems, obtaining interpretable linguistic solutions without lose of accuracy.

The idea of using F-FOCUS feature selection to get useful derived features has also been shown. This can be an interesting application to be further explored in future work.

On the other hand, we consider exploring the application of F-FOCUS to real world problems in our future research. We will try to get advantage of the use of expert knowledge for the definition of fuzzy labels, and the selection of the relevant derived features to consider.

References

[1] Hussein Almuallim and Thomas G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305, 1994.

[2] Antonio Arauzo, Jose Manuel Benitez, and Juan Luis Castro. C-focus: A continuous extension of focus. In *proceedings of the 7th online World Conference on Soft Computing in Industrial Applications*, 2002.

[3] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

[4] S. Hettich and S. D. Bay. The uci kdd archive. http://kdd.ics.uci.edu/, 1999.

[5] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994. Journal version in AIJ, available at http://citeseer.nj.nec.com/13663.html.

[6] José Jesús Castro Sánchez. *Diseño e implementación de una herramienta para la adquisición automática de conocimiento*. PhD thesis, E.T.S. de Ingeniería Informática, University of Granada, 2000.