Classification with Nominal Data Using Intuitionistic Fuzzy Sets

Eulalia Szmidt and Janusz Kacprzyk

Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01–447 Warsaw, Poland
{szmidt,kacprzyk}@ibspan.waw.pl

Abstract. The classical classification problem with nominal data is considered. First, to make the problem practically tractable, some transformation into a numerical (real) domain is performed using a frequency based analysis. Then, the use of a fuzzy sets based, and – in particular - an intuitionistic fuzzy sets based technique is proposed. To better explain the procedure proposed, the analysis is heavily based on an example. Importance of the results obtained for other areas exemplified by decision making and case based reasoning is mentioned.

1 Introduction

We deal with a classification problem, a "meta-problem" in many areas, notably computer science, decision making, etc., with nominal (categorical) data. In nominal data, names (usually belonging to a small set) are assigned to objects as labels. Initially, the only comparison possible is that if the names are the same, the two data items belong to the same category, otherwise they are different, i.e. "equality" or "inequality".

A number of approaches to classification with nominal data have been proposed (cf. Bock and Diday [3]). Basically, they boil down to some trickery to obtain a numeric assessment of nominal values, or maybe rather relations between them. Some solutions can also be found in the context of database queries and linguistic database summarization in Zadrożny [24].

For instance, in Li and Biswas [13] a similarity-based agglomerative clustering (SBAC) is proposed based on a similarity measure proposed by Goodall for biological taxonomy that gives a greater weight to uncommon feature value matches in similarity computations and makes no assumptions of the underlying distributions of the feature values. An agglomerative algorithm is used to derive a dendrogram, and then using a heuristic technique a partition of the data is extracted. Fountoukis, Bekakos and Kontos [9] present an extension of the well known decision tree approach to classification. Cheng et al. [4] propose how to define a good distance (dissimilarity) measure between patterns with nominal attributes by using adaptive dissimilarity matrices for measuring dissimilarities between nominal values. These matrices are learned via optimizing an error function on training samples. This is different than the conventionally

P. Melin et al. (Eds.): IFSA 2007, LNAI 4529, pp. 76–85, 2007.

employed value difference metric (VDM) used to define a real-valued distance measure on nominal values. De Carvalho et al. [5], [6], [7] proposed some proximity measures based on histograms. Ichino and Yaguchi [11] used a Minkowski metric, and then extended their analysis in Ichino, Yaguchi and Diday [12] to obtain a fuzzy classifier. Quinlan's [15] ID3 algorithm proved to be effective to handle both numeric and nominal data but it can be viewed to fail to handle a "topological" aspect of knowledge as it does not consider how sure the classification is, what the most typical example is, etc. To deal with these issues one has to resort to numeric analysis, notably via a similarity/proximity measure. Narazaki and Ralescu [14] proposed an alternative model which involves two stages: the configuration stage mapping the symbolic problem into a numerical domain by devising an appropriate distance measure, and then the classification of examples via the distance measure developed.

Here we propose two alternative approaches to the classification of nominal data attempting to involve merits of those approaches above using fuzzy sets (cf. Zadeh [23]), and intuitionistic fuzzy sets (cf. Atanasov [1], [2]).

2 A Brief Introduction to A-IFSs

One of the possible generalizations of a fuzzy set in X (Zadeh [23]), given by

$$A^{'} = \{< x, \mu_{A'}(x) > | x \in X\} \tag{1}$$

where $\mu_{A'}(x) \in [0,1]$ is the membership function of the fuzzy set $A^{'}$, is an A-IFS, i.e. Atanassov's intuitionistic fuzzy set, (Atanassov [1], [2]) A given by

$$A = \{< x, \mu_A(x), \nu_A(x) > | x \in X\} \tag{2}$$

where: $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ such that $0 \leq \mu_A(x) + \nu_A(x) \leq 1$, and $\mu_A(x)$, $\nu_A(x) \in [0,1]$ denote a degree of membership and a degree of non-membership of $x \in A$, respectively.

Obviously, each fuzzy set may be represented by the following A-IFS $A = \{< x, \mu_{A'}(x), 1 - \mu_{A'}(x) > | x \in X\}$. For each A-IFS in X, we will call

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \tag{3}$$

an *intuitionistic fuzzy index* (or a *hesitation margin*) of $x \in A$, and it expresses a lack of knowledge of whether x belongs to A or not (cf. Atanassov [2]). It is obvious that $0 \leq \pi_A(x) \leq 1$, for each $x \in X$.

An A-IFS gives us an additional degree of freedom, i.e. a possibility to represent more aspects of imperfect knowledge – cf. Szmidt and Kacprzyk's papers, given in the references, where applications of intutionistic fuzzy sets to group decision making, negotiations, etc. are presented.

Distances are clearly of utmost importance, and to be more specific we will use the normalized Euclidean distance between intuitionistic fuzzy sets A, B in