Approaches to Classification of Real World Datasets Using the Discernibility Concept

PhD Student
Zacharias Voulgaris
zacharias@dcs.bbk.ac.uk
www.dcs.bbk.ac.uk/~zacharias

PhD Supervisors
Dr. George Magoulas
Prof. Boris Mirkin

Project Details
Started: Jan 2005
Expected End: Sep 2008

Keywords
Classification, Pattern Recognition, Ensemble, Discernibility

Project Aims

Improve performance of classification algorithms in real world problems using the Discernibility concept.

Key details

The project introduces the Discernibility concept which is a heuristic developed for assessing how distinguishable the classes of a dataset are. The concept can be applied at the level of the individual element (denoting how distinguishable a single element of the dataset is) and at the level of the dataset (in which case it is called the Index of Discernibility). Fig. 1 shows how this concept is applied to a simple 2-dimensional dataset at both levels.

The applications of Discernibility embrace different aspects of machine learning. It has been applied successfully for upgrading the kNN classifier by creating extensions based on it [1]. Also, it has been used with success for feature selection on datasets of high dimensionality [2]. In addition, its use for the development of a data reduction technique [3] has shown promising results. Furthermore, it is researched how it can be applied for the development of a Reliability measure which can be used as a reject-option for various classifiers.

Current Work

Currently, we are working on the final stage of the project which involves the creation of an ensemble of classifiers. The classifiers are chosen to be as diverse as possible ranging from Nearest Neighbour methods to decision trees and statistical classifiers. At the same time, we are at the last stages of developing a Reliability measure based on Discernibility, as well as other measures, for building an elite of classifications which will exhibit high accuracy rate.

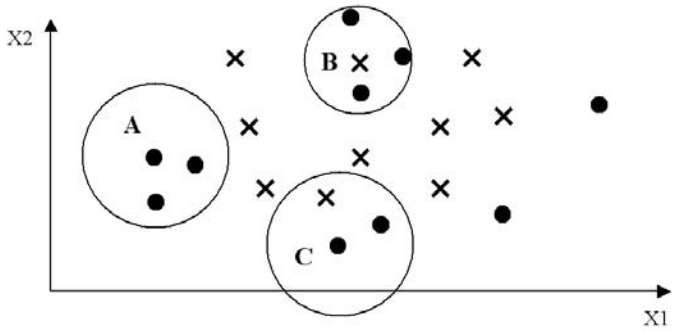

Fig. 1. Graphical representation of the Discernibility concept for three elements of a simple dataset with only two features, X1 and X2. For each pattern, the number of elements in its neighbourhood are examined and the ratio of the ones of the same class over the total number of neighbours is calculated. In this example, the discernibility of the element in the centre of circle A is ID1 = 2 / 2 = 1, that of the element in the centre of circle C is ID2 = 1 / 2 = 0.5, while that of the element in the centre of circle B is ID3 = 0 / 3 = 0. If there are 15 elements in this dataset, having ID greater or equal to 0.5, the ID of the dataset would be ID = 15 / 20 = 0.75, denoting that its classes are rather distinguishable.

Key publications

[1] Z. Voulgaris and G. D. Magoulas, "Extensions of the k nearest neighbour methods for classification problems", in *Proc. of 26th IASTED International Conference on Artificial Intelligence and Applications*, CD Proceedings ISBN: 978-0-88986-710-9, 2008, pp. 23-28.

[2] Z. Voulgaris and G. D. Magoulas, "A discernibility-based approach to feature selection for microarray data", in *4th International IEEE Conference on Intelligent Systems*, forthcoming, 2008.

[3] Z. Voulgaris and G. D. Magoulas, "Dimensionality Reduction for Feature and Pattern Selection in Classification Problems", in 3rd International Multi-Conference on Computing in the Global Information Technology, forthcoming, 2008.