Galois Connections Between Semimodules and Applications in Data Mining

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno*

Dpto. de Teoría de la Señal y de las Comunicaciones
Universidad Carlos III de Madrid
Avda. de la Universidad, 30. Leganés 28911. Spain
{fva,carmen}@tsc.uc3m.es

Abstract. In [1] a generalisation of Formal Concept Analysis was introduced with data mining applications in mind, $\mathcal{K}$-Formal Concept Analysis, where incidences take values in certain kinds of semirings, instead of the standard Boolean carrier set. A fundamental result was missing there, namely the second half of the equivalent of the main theorem of Formal Concept Analysis. In this continuation we introduce the structural lattice of such generalised contexts, providing a limited equivalent to the main theorem of $\mathcal{K}$-Formal Concept Analysis which allows to interpret the standard version as a privileged case in yet another direction. We motivate our results by providing instances of their use to analyse the confusion matrices of multiple-input multiple-output classifiers.

1 Motivation: The Exploration of Confusion Matrices with $\mathcal{K}$-Formal Concept Analysis

In pattern recognition tasks, when a classifier is provided training data in the form of feature vectors tagged with an *input pattern set* and produces for each vector a tag within an *output pattern set*, the performance of the classifier can be gleaned from the collection of pairs (g_i, m_j) of one input tag, g_i, for the input data and one output tag, m_j , produced by the classifier. These results are aggregated into a *confusion matrix*, T, whose element T_{ij} gives a "measure" of the joint event $(G = g_i, M = m_j)$, "providing an input pattern g_i to the classifier who then produces an output pattern m_j".

In the pattern recognition community we often encounter methods that use confusion matrices to analyse classification results. However, most of the times the analysis is manual and limited to the (human-based) pondering of a confusion matrix-representation like the one depicted in figure 1, where the warmer, brighter (resp. cooler, darker) colour hues are designed to be related to high occurrence (resp. to low occurrence) of events. Often, this type of analysis is used

* This work has been partially supported by two grants for "Estancias de Tecnólogos Españoles en el International Computer Science Institute, año 2006" of the Spanish Ministry of Industry and a Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2005-04264/TCM.

S.O. Kuznetsov and S. Schmidt (Eds.): ICFCA 2007, LNAI 4390, pp. 181–196, 2007.

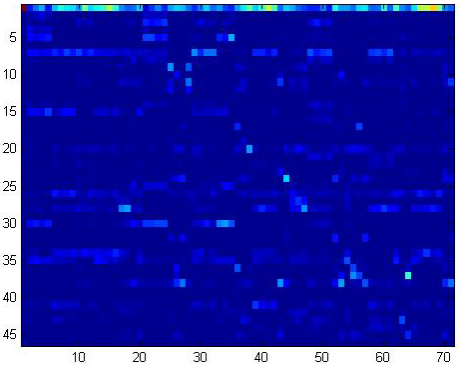

Fig. 1. Confusion matrix of the desired transformation of English phoneme labels of speech frames versus their true Mandarin phoneme labels

to bootstrap existing classifiers in order to obtain even better classification figures or simply to understand the underlying principles of the methods employed in designing the classification. In particular, in speech recognition, the designer of a system is challenged to find in this type of representation meaningful or systematic confusions to determine to what degree the behaviour of an automatic system differs from human performance.

$\mathcal{K}$-Formal Concept Analysis was introduced in [1] as a generalisation of standard Formal Concept Analysis in the sense that incidences $R \in \mathcal{K}^{n\times p}$ represented as matrices may take values in an idempotent, reflexive semifield $\mathcal{K}$ and we take $R(i,j) = \lambda$ to mean "object g_i has attribute m_j in degree λ." Adequate analogues of basic objects in Formal Concept Analysis become therefore available.

Two serious obstacles may prevent widespread adoption of $\mathcal{K}$-Formal Concept Analysis as a data exploration technique complementary to the standard theory: on the one hand, the $\mathcal{K}$-Formal Concept Analysis analogue of the main theorem of Formal Concept Analysis is incomplete and this may worry the user willing to be on a sound mathematical ground; on the other hand, [1] did not provide an algorithm for constructing the lattice of a $\mathcal{K}$-valued formal context, which prevents its use as a data-intensive exploration procedure.

In this paper, we try to explore further whether $\mathcal{K}$-Formal Concept Analysis is a proper generalisation of standard Formal Concept Analysis for finite contexts and to pave the way for the completion of the main theorem. In order to do so we introduce the *structural lattice* of a $\mathcal{K}$-Formal Context and try to relate it to the Concept Lattice of a Formal Context.

In section 2 we first review the theory of idempotent semirings and their semimodules with a view to providing the necessary objects for our discussion. In section 3.1 we present a summary of the theory of $\mathcal{K}$-Formal Concept Analysis presented in ([1], §. 3) and add a new theoretical construct, the *structural lattice* of a semimodule over an idempotent, reflexive semiring. We demonstrate in section 4 the use of this new tool to analyse confusion matrices of multiple