Text Classification Using Lattice Machine *

Hui Wang¹ and Nguyen Hung Son²

¹ School of Information and Software Engineering
University of Ulster
Newtownabbey, BT 37 0QB, N.Ireland
H.Wang@ulst.ac.uk

² Institute of Mathematics
Warsaw University
02-095, Banacha 2 Str., Warsaw Poland
son@mimuw.edu.pl

Abstract. A novel approach to supervised learning, called Lattice Machine, was proposed in [5]. In the Lattice Machine, it was assumed that data are structured as relations. In this paper we investigate the application of the Lattice Machine in the area of text classification, where textual data are unstructured. We represent a set of textual documents as a collection of Boolean feature vectors, where each vector corresponds to one document and each entry in a tuple indicates whether a particular term appears in the document. This is a common representation of textual documents. We show that using this representation, the Lattice Machine's operations are simply set theoretic operations. In particular, the lattice sum operation is simply set intersection and the ordering relationship is simply set inclusion. Experiments show that the Lattice Machine, under this configuration, is quite competitive with state-of-the-art learning algorithms for text classification.

1 Introduction

Lattice machine, proposed in [5], is a general framework for supervised learning. Two of its components can be specified to suit different situations. In [5] the Lattice Machine was presented to work on structured data, which is in the form of database relations. In this paper we re-configure the Lattice Machine so that it works on unstructured textual data. We will show that adopting the common Boolean feature vector representation of textual documents, two of the Lattice Machine's components – ordering relation and sum operation – are simply set inclusion and set intersection respectively. This approach will be validated by

* The authors gratefully acknowledge support by the KBN/British Council Grant No WAR/992/151
experimentation using benchmark (textual) datasets, and it will also be compared with similar approaches.

In the rest of the paper, we first of all present a brief review of the Lattice Machine. Then we specify the Lattice Machine for use with unstructured textual data. We present a detailed example to illustrate the configured Lattice Machine, along with experimental results. Finally we summarise and conclude the paper.

2 A brief review of the Lattice Machine

Given a dataset represented as a database relation, a tuple is a vector of attribute values, a hyper tuple is a vector of sets of attribute values, and a hyper relation is a set of hyper tuples. The notion of hyper tuple is a generalisation of database tuple. Examples of a simple relation and a hyper relation are shown in Tables 1 and 2 respectively.

It has been shown [5] that an elegant structure is implied among hyper tuples - the collection of all hyper tuples in a given domain is a semilattice under the following ordering relation $\leq$:

$$ hyper_{tuple1} \leq hyper_{tuple2} \iff hyper_{tuple1}(A) \subseteq hyper_{tuple2}(A), $$

for all $A \in U$ and $U$ is the set of attributes. We call this domain lattice.

In a domain lattice, a labelled dataset (training data) corresponds to a labelling of the lattice. For example, the following labelled dataset corresponds to the labelled lattice in Figure 1.

- $E$: $\langle$Large, Red, Triangle, $+$\rangle
- $G$: $\langle$Large, Blue, Circle, $+$\rangle
- $J$: $\langle$Small, Blue, Triangle, $-$\rangle
- $L$: $\langle$Small, Green, Rectangular, $-$\rangle

Given a labelled lattice, a straightforward representation of the labelling is by sets - each set consisting of all data units with the same label, and a simple classification is by enumeration - if a new data unit is the same as one element in the set representation, then it is classified by the label of this element; otherwise, it is marked as unknown. This is in fact the idea of rote learning. Clearly there is no generalisation in this kind of learning. The nearest neighbour method goes one step further. In its basic form, the representation is also by sets, but the classification is based on some distance measure. There is generalisation in this kind of learning, but distance measures can be troublesome in cases where discrete attributes are involved.