Probabilistic Relational Modelling of Mammographic Images[1]

Nivea Ferreira Peter J.F. Lucas

Institute for Computing and Information Sciences,
Radboud University Nijmegen
{nivea,peterl}@cs.ru.nl

Abstract

1 Introduction

Screening mammography is a breast examination which uses X-ray imaging to aid in the early detection and diagnosis of breast abnormalities in asymptomatic women. Early detection of breast cancer from mammograms is of crucial importance to improve the prognosis of patients with breast cancer. However, it presents a number of challenges, specially in the areas of data mining and machine learning. In such context, the main aim is the development of computer-aided detection (CAD) systems to assist radiologists in the reading and interpretation of exams.

Mammograms are typically formed by different projections, or views, of each of the patient's breasts, being mediolateral oblique (MLO) and craniocaudal (CC) the most common ones. As a consequence, an appropriate interpretation of mammograms require that information across the views are correlated. Besides, it is not unusual (in such domain) to come across unbalanced datasets, containing a very small percentage of true cancers. Moreover, the common assumption of learning algorithms that data is independent and identically distributed might not be applicable. In fact, in a dataset consisting of information from detected regions in a certain breast, it is most commonly the case that such regions would be related to one another. For instance, a suspicious region in MLO view may have a corresponding region in CC view.

The aim of the present work is, firstly, to show that object-oriented database theory offers a natural start for the design of pattern recognition techniques in the breast cancer domain and, secondly, to explore the use of relational probabilistic methods for the detection of breast cancer in the screening mammography domain.

2 Probabilistic Relational Trees

In relational domains the data instances are no longer recorded in homogeneous structures as commonly used in machine learning. Rather, the data is organised in terms of objects that have different attributes, and are linked to one another. The estimation of probability distributions for relational probabilistic models does not automatically assume that instances are independent and identically distributed, the almost standard assumption of maximum likelihood estimators.

Probabilistic relational models define a generic dependency structure at the level of item types (in contrast to defining the dependency structure over attributes of specific objects). Typing items, and parameters across items of the same type, enables generalisation from a single instance by decomposing the data graph into multiple examples of each item type and building a joint model of dependencies between, and among, attributes of each type [1].

[1]The full version of this paper is published at the *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, 2009*.

Relational probability trees (RPTs) [2] build a model which shows a selective and intuitive representation of domain knowledge. Its learning algorithm takes a set of subgraphs as input, where each subgraph contains a target object to be classified and a set of (other) objects which form this target object's relational neighbourhood. It then constructs a probability estimation tree to predict the target label given i) the attributes described for the target object; ii) the attributes of objects, as well as links, in the target object's neighbourhood; and, iii) aggregated attributes and links in such neighbourhood. The algorithm searches over the space of binary relational features in order to obtain a split of the data, taking into account feature scores and correlation among features.

For instance, consider the segment of an RPT shown in Figure 1. In this case, we are building a model for the classification of MLO regions. The first node shown checks the value of attribute *score* (of suspiciousness) of associated CC regions: if this value is greater than 1.62 in at least four CC regions, then we proceed to the subtree on the right; otherwise, to the one on the left. At the subtree on the right, the attribute *spicul* (spiculation) of the MLO region is tested: if greater or equal than 1.05, with probability 0.57 the given region is classified as a non-cancerous region; otherwise, the region is cancerous with probability 0.84. The reasoning is similar to at the subtree on the left, however here another attribute, distance to skin, is tested. In short, on classifying MLO regions the attributes of related CC regions are also considered.

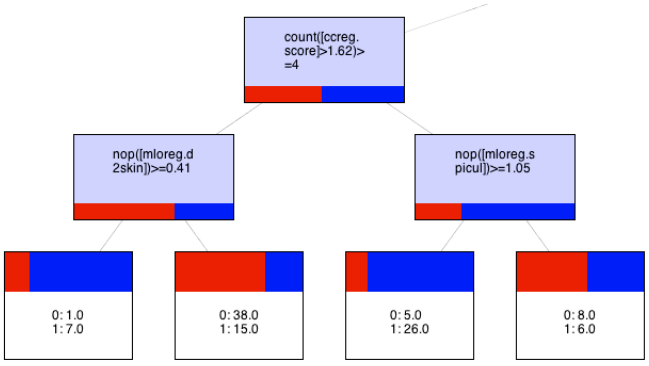

Figure 1: A fragment of a relational probability tree for MLO regions.

3 Breast Cancer Modelling

Different models were build using different sets of features and different depth for the relational probability trees. Despite the different settings, accuracy of models were usually high. This is due to the fact that the distribution of the data used (in terms of non-cancerous and cancerous regions) is highly unbalanced. However, ROC and AUC analysis, allow us to state that learned models were compatible to previously developed models of the domain based on the same data. This encourage us to continue exploring relational models for analysing mammograms.

We are, for instance, able to show that RPT models can be used in order to improve the predictions made by a previously developed CAD system, begin valuable not only for its prediction power but for its intelligibility. We consider the use of relational probabilistic models for the breast cancer domain as a natural and comprehensible way of representing the various domain entities, as well as the uncertain relations among those. Besides, the richness of the domain itself empower the conceptual modelling of data, and learning of relational models given the considerable amount of data available.

References

[1] J. Neville and D. Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 2007.

[2] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, 2003.