Data Reduction for Supervised Learning in Medical Image Analysis

Introduction of a Measurement Decision Tree

Armin Stoll[1], Andrea Fränzle[1], Rolf Bendl[1,2]

[1]Medical Physics in Radiation Oncology, DKFZ Heidelberg
[2]Medical Informatics, Heilbronn University
a.stoll@dkfz-heidelberg.de

Abstract. Segmentation of organs at risk for radiation therapy planning is a time-consuming task. Knowledge-based segmentation is in the focus of research to apply semi-automatic segmentation without user-interaction. It turns out that only integrating general knowledge into a knowledge base is not appropriate to deal with complex anatomical conditions among humans. Supervised learning is therefore investigated to provide individually given knowledge and adapt general knowledge to the individual case. Since deriving multi-dimensional feature spaces in medical images can cause huge amount of redundant data, a measurement decision tree is presented and its data reduction performance evaluated.

1 Introduction

Segmentation of organs at risk (OAR) for radiation therapy planning is a time-consuming and complex task. It is often done manually or semi-automatic. Volumes of interest (VOI) and their spacial location play a major role in radiation therapy planning. It enables to quantitatively simulate radiation therapy under certain therapy conditions. New developments in radiation therapy planning brought new requirements to the segmentation task on medical images. Automatic segmentation would allow to repeat treatment planning more often since this time-consuming task is still a limiting factor. Many semi-automatic segmentation algorithms are already developed. One drawback is that complex configuration can remain necessary to achieve expected segmentation results. As a consequence, semi-automatic segmentation also turns into a knowledge intensive task. Knowledge-based systems are in the focus of research to provide relevant knowledge. It is intended to trigger semi-automatic segmentation without user interactions [1]. But, it turns out that only incorporating general knowledge is not sufficient to deal with complex anatomical conditions among humans. Pattern recognition methods are therefore investigated to acquire individually given knowledge. As a result, semi-automatic segmentation algorithms are triggered by individually adapted knowledge.

This contribution introduces a measurement decision tree (MDT) for supervised learning purposes. It is intended to reduce the number of measurements for the training of classifiers on large scale medical image data.

2 Material and Methods

Knowledge-based segmentation [2, 3, 4] in radiation therapy planning is mainly investigated for computed tomography (CT) since it is the principal foundation for physical radiation dose simulation. CT images show X-ray attenuation coefficients in Hounsfield Units [HU] within a tiny volume – a voxel. To classify single transversal slices into body regions, different classifiers are already investigated [1, 5]. Further investigations, now focusing on multi-dimensional image processing and voxel classification, easily exceed the amount of training data by a factor of 2 to 10. This is due to statistical features (measurements) derived from small subregion for each voxel. For instance, minimum, maximum, median, variance and Haralick features [6]. In the same context, labels are derived from already existing manual segmentation (fig. 1). The amount of training data can easily turn into main memory requirements exceeding the 3 GB threshold on 32 bit computer systems. It is observed that among all measurements many are redundant represented (e.g. a lot of background measurements are fixed to -1024 HU and therefore highly over-expressed). Cross-validation on redundant data becomes very inefficient and turns into superfluous computational efforts.

3 Results

Decision trees are well known from the context of machine learning [2]. Unlike other classifiers, models produced by decision trees are easy to interpret. The reasoning process is transparent and straightforward from the root node to its leaf nodes. In similar prospect a MDT is proposed to arrange measurements m and its labels l in a tree representation to reduce redundancies.

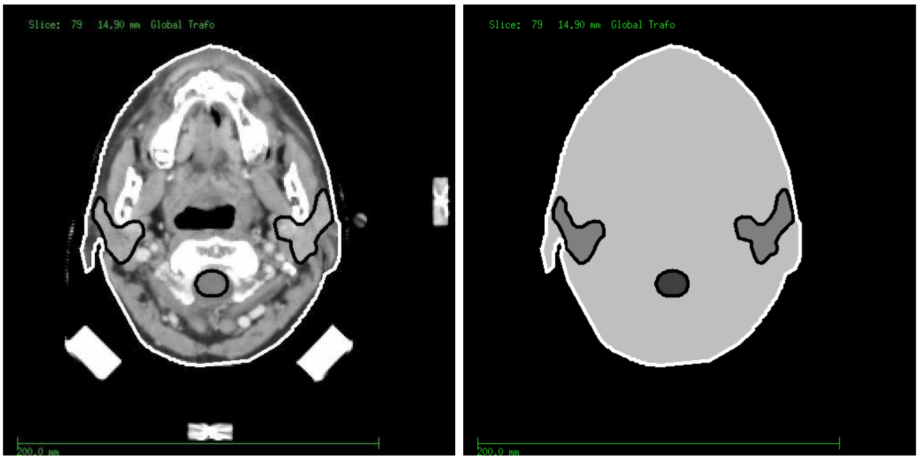
Slice: 79 14.90 mm Global Trafo
200.0 mm
Slice: 79 14.90 mm Global Trafo
200.0 mm

Fig. 1. Training data. Left: manually segmented contours of a head&neck case. Right: derived labels per voxel (background, human anatomy, spinal cord and parotids).

3.1 Measurement Decision Tree

Each measurement m consists of several features x_i. Each feature is represented on a separate level in the MDT (fig. 2). Two types of nodes are known: feature nodes and label nodes. A feature node (FN) consists of value x_i (e.g. a numerical value, character or text) and a list of subsequent feature nodes with values x_{i+1}. A label node (LN) is always a leaf node and contains a label l and a counter variable c. In Python syntax the implementation is the following:

```
class FeatureNode:
  def __init__(self, value, children=[]):
    self.value = value
    self.children = children
  def __eq__(self, value):
    return self.value == value

class LabelNode:
  def __init__(self, label, counter=1):
    self.label = label
    self.counter = counter
  def __eq__(self, label):
    return self.label == label
```

Training To build a MDT, a root node (RN) has to be created first. The first feature x_1 of the first measurement m_1 is attached to the RN as FN. The second feature x_2 is attached to the previous FN. This process is continued until every feature of the first measurement is integrated as the first branch in the MDT. Finally, a LN is attached to the last FN and its counter variable is initialized with value 1. Insertion of the second measurement m_2 begins at the RN again.

```
1 1:0 2:0 3:3.0
1 1:0 2:0 3:3.0
2 1:0 2:0 3:3.0
1 1:-100 2:0 3:3.0
1 1:-100 2:0 3:4.0
3 1:-1024 2:0 3:2.0
3 1:-1024 2:0 3:2.0
3 1:-1024 2:0 3:2.0
2 1:-1024 2:4 3:2.0
2 1:-1024 2:4 3:2.0
```

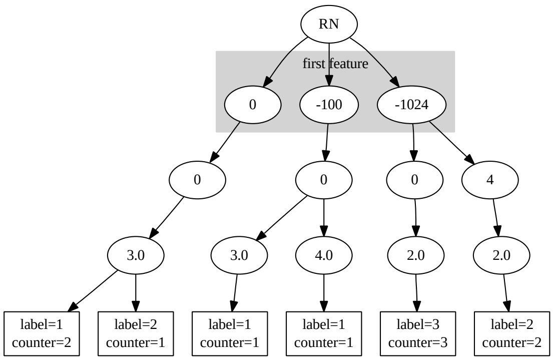

Fig. 2. Example: left some particular training data (format: label, feature1:value1, feature2:value2 etc.) and right the resulting MDT.

If the first feature is already present, no new node is created. If not, then a new FN is appended to the current node. This work flow is true for every feature x_i. The counter variable of a LN is either incremented by one if a corresponding LN already exists or a new one created. A simple example is shown in fig. 2. The implementation of the *insert* function is given as the following:

```
def insert(measurement, label):
  node = root
  for m in measurement:
    if m in node.children:
      i = node.children.index(m)
      node = node.children[i]
    else:
      node.children.append(Node(m,[]))
      i = node.children.index(m)
      node = node.children[i]

  if label in node.children:
    i = node.children.index(label)
    node.children[i].value += 1
  else:
    node.children.append(LeafNode(label,1))
```

As an advantage, no information get lost when measurements are represented as MDT. If more than one label is assigned to a measurement, different decision strategies become possible. For instance, conserving the measurement for every label, only for the label with maximum occurrence, for all labels that exceed a certain threshold or that one that gains absolute majority.

This MDT was applied to reduce the measurement space for five head&neck cases. Six features were derived within three-dimensional CT data for every voxel. The total number of measurements per case is depicted in tab. 1. The training data is reduced by integrating each measurement into a MDT and transforming it back into a linear sequence. Each label is now non-redundant represented by exactly one measurement. The total number of remaining measurements and its reduction is also shown in tab. 1. Finally, all reduced training sets are merged into one single training set that comprises all cases. The total number of remaining measurements is $6\ 285 \cdot 10^3$. As a result, measurements from different classes are now more balanced distributed and redundancies strongly reduced.

4 Discussion

The proposed MDT allows to reduce redundant measurements on large scale medical image data to $\approx 7\,\%$ of its original size. Measurements are represented exactly once in the feature space. Adverse side effects on single classifiers (e.g. k-nearest neighbours classifier) have to be individually evaluated. The main benefit

Table 1. Evaluation: measurement reduction performance by insertion into the MDT.

Case	# Measurements	# Measurements remaining	Reduction [%]
1	$40 \cdot 10^6$	$1\ 668 \cdot 10^3$	96
2	$21 \cdot 10^6$	$862 \cdot 10^3$	96
3	$28 \cdot 10^6$	$1\ 151 \cdot 10^3$	96
4	$25 \cdot 10^6$	$1\ 018 \cdot 10^3$	96
5	$41 \cdot 10^6$	$6\ 381 \cdot 10^3$	84
Mean:	$31 \cdot 10^6$	$2\ 216 \cdot 10^3$	93

results in the reduction of computational effort for supervised learning and cross-validation since the chance of testing redundant data reduces tremendously. As a result, individual knowledge aquisition on voxel scale becomes possible for knowledge-based segmentation purposes [2, 4].

References

1. Stoll A, Bendl R. Wissensakquisition mit methoden der mustererkennung zur wissensbasierten segmentierung von risikoorganen in CT-Bilddaten. Proc BVM. 2008; p. 41–5.
2. Schmidt G, Kietzmann M, Kim J, et al. Cognition network technology for fully automatic 3D segmentation of lymph nodes in CT data. In: Dössel O, Schlegel WC, editors. World Congr Med Phys Biomed Eng. vol. 25/IV; 2009. p. 1365–7.
3. Foruzan AH, Zoroofi RA, Hori M, et al. A knowledge-based technique for liver segmentation in CT data. Comput Med Imaging Graph. 2009;33(8):567–87.
4. Klinder T, Wolz R, Lorenz C, et al. Spine segmentation using articulated shape models. In: Proc MICCAI; 2008. p. 227–34.
5. Fränzle A, Stoll A, Bendl R. Ermittlung einer kranial-kaudalen Korrespondenz in MR-Aufnahmen. Proc BVM. 2009; p. 232–6.
6. Medical image analysis of 3D CT images based on extension of Haralick texture features. Comput Med Imaging Graph. 2008;32(6):513–20.