Retrieval with Clustering in a Case-Based Reasoning System for Radiotherapy Treatment Planning

Gulmira Khussainova, Sanja Petrovic, Rupa Jagannathan

Operations Management and Information System Division, Nottingham University Business School, University of Nottingham, Nottingham NGS 1BB, UK
E-mail: psxgk@nottingham.ac.uk, sanja.petrovic@nottingham.ac.uk, rupa.jagannathan@nottingham.ac.uk

Abstract. Radiotherapy treatment planning aims to deliver a sufficient radiation dose to cancerous tumour cells while sparing healthy organs in the tumour surrounding area. This is a trial and error process highly dependent on the medical staff’s experience and knowledge. Case-Based Reasoning (CBR) is an artificial intelligence tool that uses past experiences to solve new problems. A CBR system has been developed to facilitate radiotherapy treatment planning for brain cancer. Given a new patient case the existing CBR system retrieves a similar case from an archive of successfully treated patient cases with the suggested treatment plan. The next step requires adaptation of the retrieved treatment plan to meet the specific demands of the new case. The CBR system was tested by medical physicists for the new patient cases. It was discovered that some of the retrieved cases were not suitable and could not be adapted for the new cases. This motivated us to revise the retrieval mechanism of the existing CBR system by adding a clustering stage that clusters cases based on their tumour positions. A number of well-known clustering methods were investigated and employed in the retrieval mechanism. Results using real world brain cancer patient cases have shown that the success rate of the new CBR retrieval is higher than that of the original system.

1. Introduction

Every day in England 11 people are diagnosed with brain cancer and it claims 9 lives [1]. The process from diagnosis to treatment contains many steps. Medical images such as computed tomography (CT) and magnetic resonance imaging (MRI) are taken to detect the location of the tumour. After the tumour and the organs surrounding the tumour, so called organs-at-risk (OAR) are outlined the treatment planning process starts. In radiotherapy treatment planning the aim is to deliver sufficient radiation to kill tumour cells while trying to spare healthy tissues in the tumour surrounding area.

Radiotherapy treatment planning is usually a trial and error process and needs the expertise of one or more experienced medical physicists and oncologists. It can take up to 3-4 days depending on the complexity of the plan. Developing a treatment plan depends highly on the expertise and experience of the medical staff. Most software applications that are in use to facilitate radiotherapy treatment planning are based on optimisation methods [2, 3, 4, 5]. However the drawbacks of such methods is that they do not utilise the experience of the medical staff, which can provide valuable information when designing a treatment plan for an individual patient.

Case-based reasoning (CBR) uses previous experience to understand and solve new problems based on similar problems solved in the past [6]. When creating a treatment plan for a new
patient medical physicists try to retrieve/remember similar cases that they have encountered in the past along with their treatment plans and try to map it to the new case. Therefore, using CBR seems to be a natural decision in developing an application that will help medical staff to design a treatment plan. The purpose of an automated software application system is to save medical physicists’ time. In addition, such system could be used to train junior medical staff.

The research presented in this paper builds on the CBR system for 3D conformal radiotherapy treatment planning of brain cancer which was developed in collaboration with the Nottingham University Hospitals Trust, NHS, Nottingham City Hospital Campus and is described in Jagannathan et.al. [7]. Each case in the case base consists of the case attributes that describe the relationship between the tumour and OARs of that patient and the treatment plan that was used to treat the patient. The treatment plan consists of the number of radiation beams and beam angles. Given a new patient case the CBR system computes the similarity between the attributes of the new case and the cases in the case base and returns the most similar case along with its treatment plan. This treatment plan can then be adapted by medical physicists to meet the specific requirements of the new patient. Adaptation of the retrieved solution is a crucial part of the CBR process that ensures that the solution can be applied to the new case.

The CBR system was tested by medical physicists on new patient cases. However, the results showed that the treatment plans of some of the retrieved cases could not be adapted to be applied to the new case as their tumour position was very different to the tumour position of the new case resulting in highly unsuitable beam angles.

Treatment planning aims to deliver the maximum of the prescribed radiation on the centre of the tumour. Therefore, beams placement is largely determined based on the position of the tumour and points where beams intersect should be placed at the centre of the tumour.

In this paper, we revised the existing CBR retrieval mechanism to retrieve cases with more suitable tumour positions. It is not expected to find a case in the case base with the tumour positioned in the exact same location as the new case. The aim is to consider only the cases whose tumour positions is very close to the new case so that beam angles in the retrieved treatment plan can be adapted. For this purpose cases are clustered by their tumour positions so that during retrieval the system considers only the cases that are in the same cluster as the new case. The results of the revised CBR system are compared with the previously developed CBR system.

The continuation of this paper is as follows. Section 2 contains a description of the developed CBR system and the insights learned from the CBR validation. Section 3 describes methods implemented for determining the optimal number of clusters for tumour position and clustering itself. Section 4 presents the computational results of clustering of cases in the current case base. Section 5 compares the performance of the original and revised CBR systems. The paper finishes with the conclusion in section 6.

2. Background and related work

This section will introduce the existing CBR retrieval system, the insights learned from its use by medical physicists and the required revisions to the existing retrieval mechanism to improve the suitability of the retrieved treatment plan to the new case.

2.1. Existing CBR system

The CBR system has been implemented to facilitate in developing a radiotherapy treatment plan for brain cancer patients. It contains a case base of 86 brain cancer patient records received from Nottingham City Hospital containing descriptions of patient cases and their treatment plans.

In the Nottingham City Hospital, after examining the image files of patients the planning treatment volume (PTV) containing the tumour and organs at risk (OAR) are outlined in the treatment planning system by the oncologists. The following OAR are considered in the
treatment plan: the right and left lens of the eyes, right and left optical nerves, the spinal cord, optic chiasm and brainstem. The treatment plan is generated by medical physicists in a trial-and-error fashion by placing beams on the patient image in the treatment planning system to achieve an acceptable dose of irradiation.

In the existing CBR system the case most similar to a new case is retrieved using a similarity measure based on the spatial relationship between the tumour and OAR, the volume of the tumour and the ratio of the tumour to the patient body. Not all of the OAR mentioned above are present in all cases. Depending on the tumour location some OAR are absent as they are not affected by radiation beams. For example, if the tumour is located in the front part of the head then the brainstem might not be an OAR as no risk is posed to it due to its large distance to the tumour. First, filtering of the case base is performed, which considers for retrieval only cases with the same OARs as the new case. Then the similarity between the new case and the filtered cases in the case base is computed using a weighted nearest neighbour (wNN) similarity measure with local weights [8]. The retrieval consists of two phases. First, the retrieved case suggests the number of beams, and then among the remaining filtered cases with this number of beams the most similar one suggests the beam angles.

In consultation with medical physicists the following case attributes that affect the treatment plan configuration were identified:

V - volume of the tumour (given in mm).
A - angle between the line connecting the origin of the coordinate system with the centre of the tumour and with the centre of OAR (0-360).
E - minimum distance between the edge of the tumour outline and OAR outline (mm).
R - the ratio of the tumour volume to the volume of patient’s body (Body - Tumour volume ratio).
Dt - minimum distance between the body edge and the tumour outline (Body - Tumour volume distance) (mm).
P - the relative position of the tumour to OAR (Tumour - OAR spatial relationship). It contains six labels: left, right, posterior, anterior, superior and inferior.

The treatment plan of the retrieved case is applied to the new patient. At the current stage of development, the CBR system includes two treatment plan parameters: the number of radiation beams which are applied to the tumour and the angles of those beams. Beam angles are defined by the gantry angle, patient couch angle and collimator angles.

The treatment plans in the case base were generated by medical physicists and validated by the oncologists. Our aim is to mimic the way medical physicists generate treatment plans, and at this stage of the project we use the plans generated by them to evaluate our CBR system. We employed the leave-one-out cross validation technique to determine the accuracy of the retrieval system and the similarity measure by repeatedly removing a case with a known solution from the case base, so called target case, retrieving the most similar case from the remaining cases in the case base and comparing the solution of the retrieved case with the actual solution of the target case. The difference between two solution forms the retrieval error. The average error obtained by making each case in this fashion consecutively the target case gives an indication of the performance of the similarity measure. This method is known as the leave-one-out cross validation technique. The error is calculated in terms of the difference between the beam angle(BA) error and the beam number(BN) error in the treatment plans of the retrieved and target cases. The BA error is calculated based on the difference between the beam angles and the BN error, by the difference in beam numbers between the treatment plans of the retrieved and the target cases. The aim is to retrieve a treatment plan with BN error=0 and minimal BA error.
2.2. CBR system validation

The developed CBR system was presented to medical physicists who used it for the treatment planning of 6 new patients. This section describes the insights learned from this validation.

In all 6 cases, the system retrieved treatment plans with the same beam number as the new case. For 2 out of the 6 new patient cases the retrieved cases provided treatment plans that were a good starting point for adaptation; for 2 cases the retrieved cases suggested treatment plans with some acceptable beam angles and some beam angles that medical physicists would not use; finally, for the last 2 new cases the suggested beam angles could neither be used nor adapted for the new case.

When the beams are placed the radiation is focussed at the point, known as the isocentre, where all beams meet and ideally this point should be placed at the centre of the tumour. In all cases in our case base the tumour position in the patient case and the isocentre positions in the treatment plan coincide.

When adapting cases medical physicists noted that the isocentre of the retrieved treatment plan was often misplaced with respect to the tumour position centre of the new patient. The tumour position relative to OAR is captured in the similarity measure by attribute $P$ described earlier. However, this attribute was not sufficient to define the tumour position as the isocentre of the suggested beam in the retrieved treatment plan did not match the tumour centre of the new case.

The tumour position is important as it defines how and from which side of the head beams will enter, e.g. if the tumour is located in the left side of the head then the beams are placed to enter the head from the left side and having beams enter from the right side would mean passing through a lot of healthy tissues. In the cases where the retrieved beams are misplaced in such a manner it is easier to design a new treatment plan from scratch than to adapt the retrieved treatment plan. Therefore, in order to avoid retrieving such non-adaptable treatment plans, it was decided to take into account the tumour position of cases during the CBR retrieval system.

The evaluation of the retrieved treatment plans for the six new cases showed that that metrics used to compute the retrieval error in terms of BA and BN do not give an indication of how suitable the retrieved plan is for adaptation. It is possible that the retrieved treatment plan is difficult to adapt or unsuitable for adaptation in spite of having a small retrieval error. For this reason, the authors developed an adaptation-guided retrieval mechanism, where the modified CBR system retrieves a case that is not only the most similar to the new case but also can be adapted to meet the requirements of the new patient. The results of this work are in the process of publication.

2.3. Tumour position and isocentre for the retrieval mechanism

The aim is to match the isocentre of the retrieved treatment plan with the tumour position of the target case. If the tumour positions of the retrieved and target cases match then the isocentre of the retrieved treatment plan should be a good match for the target case. After the OAR filtering, we filter cases based on their tumour positions. However, tumour positions usually do not match exactly in different cases. Thus, we look for a case to retrieve with a tumour position located close to the tumour position of the new case. All cases in the case base are clustered based on their tumour position. Given a target case, the system finds all cases in the case base with the same tumour position cluster as the target case. Only these cases are available for the retrieval. The system then computes the similarity between the target case and these cases, and retrieves the most similar case.

3. Clustering methods

We use the k-means clustering algorithm [9] to find the clusters of cases based on their tumour positions. K-means is a classical robust clustering technique used when a number of clusters
in the dataset is fixed a priori [10]. The k-means clustering algorithm divides M data samples in N dimension into k clusters so that the average squared distance from data points to the cluster centre, referred to as mean, is minimal. The k-means algorithm is iterative and in every iteration each data point is assigned to the closest cluster mean and then the means are recalculated. The optimal solution is found when the assignments no longer change the distance between data points and the cluster mean.

In the anatomical coordinate system that is used in radiotherapy treatment planning the X coordinate increases from the left side of the body towards right, and thus, it defines left and right position of the tumour, the Y coordinate increases towards the anterior part of the body and defines anterior and posterior position of the tumour and the Z coordinate increases towards the superior part of the body and therefore determines superior and inferior position of the tumour. The tumour position in the head/brain described by 3 coordinates: X, Y and Z is shown in Figure 1.

![Anatomical Coordinate System](image)

**Figure 1.** The head and brain anatomical coordinate system [11, 12].

We need to find the number of clusters to use with our data set first. Based on the coordinate system described in Figure 1 we can identify 8 regions in the brain: (Left, posterior, superior), (Left, posterior, inferior), (Right, posterior, inferior), (Right, posterior, superior), (Left, anterior, inferior), (Right, anterior, superior), (Left, anterior, superior), (Right, anterior, inferior).

Based on this we can cluster our data set into 8 clusters. However, our data set might not be distributed evenly and in some clusters there might not be any cases or cases can fall into just small number of clusters.

Therefore, it is required to find the optimal number of clusters based on our data set. In order to define the number of clusters we make use of off-the-shelf techniques available in R packages. R is an open source computing environment for statistical analysis of a data set.

There are many methods to suggest the optimal number of clusters in the data set. Some of them are designed for large data sets and are therefore, not suitable for our case base. To find the optimal number of clusters we applied 3 clustering methods to the data set. In particular, first we used a heuristic clustering method, known as partitioning around medoids (PAM) [13]. They partition data into clusters in such a way as to minimise the distance between points labelled to be in a cluster and a point designated as the center of that cluster. Heuristic methods look for a partitioning that is good enough for a given data set. Second, we used model-based clustering method which belongs to the class of probability based methods. Probability models analyse the data set and make assumptions about clusters. Both of these 2 methods described look for normally distributed clusters. Thus, finally, we used Calinski-Harabasz clustering [15], which does not look for normally distributed clusters compared to the previous methods.

4. Clustering computational results
This section compares the results of methods for determining optimal number of case clusters.
4.1. PAM clustering results

The PAM method suggested 2 clusters based on tumour position data illustrated in Figure 2. Component 1 and 2 are principal components which define the distribution of the data. The number of components is equal to the number of found correlated data points.

![Figure 2. Cases clustered with PAM method](image)

PAM provides a novel graphical display, the silhouette plot, which allows the user to select the optimal number of clusters. Silhouette plot in Figure 3 shows the clustering measure for each case and how it fits into the assigned cluster. This is done by comparing how close the case to other cases in its own cluster and cases in other clusters. Each line on the silhouette plot corresponds to a case. The Y axis represent cases and X axis shows how well the cases
fit the assigned cluster. For example, values near 1 mean that the corresponding case fits well into the cluster and values near 0 mean that the case could belong to another cluster. In each cluster the value is displayed from smallest to largest. The information in the silhouette plot is used to determine the optimal number of clusters. For the current case base the silhouette plot suggests 2 clusters. We can see that most of the cases belong to a cluster 1, and a smaller number of cases belong to cluster 2. The average silhouette width of 0.9 suggests that 90% of cases clustered were fit well by the suggested method.

4.2. Model-based clustering results
Model-based clustering with Bayesian probability clustered cases into 10 different probability models.

![Figure 4. Result of Model Based Clustering](image)

Figure 4 shows that all models converge to a point which suggests 3 to be the optimal number of clusters. The X axis represent the number of clusters and the Y axis their distribution in the space based on Bayesian probability.

4.3. Calinski-Harabasz clustering results
The Calinski-Harabasz clustering analysis presented in Figure 5 suggests 10 clusters to be the optimal solution. In Figure 5 the X axis represents the number of cases and the Y axis represents the number of clusters. Different shades represent different case clusters. The tumour position data of all cases in the case base grouped in 10 clusters can be seen from Figure 6.
Figure 5. Calinski-Harabasz clustering analysis

Figure 6. Cases clustered into 10 clusters as suggested by Calinski-Harabasz

The suggested number of clusters from the three methods above, 2, 3, and 10, were used by k-means algorithm to cluster cases by their tumour positions. We also consider clustering into
8 predefined clusters based on the geometry of a brain anatomical coordinate system given in Figure 1.

5. Revised CBR retrieval results and discussion
Experiments were carried out to assess the clustering of cases based on their tumour positions and number of clusters on the retrieval mechanism performance. Experiments were run using 2, 3, 8 and 10 clusters. Leave-one-out cross validation method was used to measure the success rate of the CBR system.

The Euclidean distance between two isocentre points \((X_1, Y_1, Z_1)\) and \((X_2, Y_2, Z_2)\) is calculated using the formula below:

\[
d(isocentre_1, isocentre_2) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2}
\] (1)

Table 1 presents the comparison between the original and revised CBR systems. In addition to BN and BA errors we also measure the distance between the isocentre of the retrieved treatment plan and known original treatment plan of the target case.

Table 1. Comparison between retrieval results of the original CBR system and the revised CBR system with clustering

| CBR system     | Average isocentre distance (mm) | Average BN error | Average BA error | Success rate change w.r.t BN% | Success rate change w.r.t BA% |
|----------------|---------------------------------|------------------|-----------------|------------------------------|------------------------------|
| Original CBR   | 207.87                          | 0.4              | 24.36           | -                            | -                            |
| CBR with 2 clusters | 107.5                         | 0.4              | 24.1            | 0                            | 1.06                          |
| CBR with 3 clusters | 108.6                         | 0.4              | 24.4            | 0                            | -0.16                         |
| CBR with 10 clusters | 30.29                        | 0.28             | 21.90           | 30                           | 13.79                         |
| CBR with 8 clusters | 31.26                         | 0.35             | 26.99           | 12.5                         | -10.79                        |

We define the CBR system’s success rate by the beam angle and beam number errors, thus, the improvement of these values over all cases in the case base denotes the overall improvement of the revised CBR system’s performance. That is, the success rate is defined by the percentage of the improvement in average BA and BN error values in the revised CBR system compared to the ones in the original CBR system.

The revised CBR system retrieved treatment plans with decreased isocentre distances between the retrieved treatment plan and known treatment plan of the target case compared to the original CBR system.

The average BA error value has decreased in CBR system with 2 clusters. However, in the CBR system with 3 clusters the BA error has increased by 0.16% compared to the average BA error in the original system. The average BN error value remains the same as in the original CBR system, 0.4, for the revised systems with 2 and 3 clusters. The revised CBR systems with 2 and 3 clusters retrieved the same cases as the original CBR system in 85% of cases. Although the revised CBR system with 3 clusters had better isocentre than the original CBR system the average BA error is worse than the one in the original system.

The revised CBR system with 8 clusters has decreased isocentre distances between retrieved case and target case from 207.87, in the original system, to 31.26 in the revised system. The average BN error has improved by 12.5% from 0.4 to 0.35. However, the BA error value has
increased compared to the original CBR system from 24.36 to 26.99. Thus, the lower BN error in the revised CBR system with 8 clusters was obtained at the expense of BA error. That is, cases retrieved treatment plans that suggested better solution with respect to beam number, however, worse solution with respect to beam angles compared to the original CBR system. Thus, the success rate of the revised CBR system is lower by 10.79% with respect to BA error. The revised CBR system with 8 clusters retrieved cases different from the original CBR system in 71% of cases.

The CBR system with 10 clusters performed the best and retrieved cases with the isocentre distance decreased to an average of 30.29. Average BA error value has decreased by 13.79% from 24.36 to 21.0. The average BN error has decreased by 30% from 0.4 to 0.28. The revised CBR system with 10 clusters retrieved cases different to the original CBR system in 80% of cases.

To summarise, the revised CBR system using clustering of cases retrieves cases whose treatment plans have isocentres that match the isocentre of the target cases better than the original system. Although, all of the revised CBR systems have improved isocentre distances compared to the original CBR system, the beam angle error between the retrieved and target cases did not improve in all of them. The improvement in beam angle error was noted in the CBR system with 2 clusters. The CBR system with 3 clusters did not improve the original CBR system. In the CBR system with 8 clusters improvement in beam number error was noted at the expense of beam angle error. The revised CBR system with 10 clusters performed best with respect to both the average beam angle and beam number error. Therefore, we determine 10 clusters to be an optimal number for our data set and use the CBR system with 10 clustering.

6. Conclusion

The revised CBR system with 10 clusters performed best with respect to both the average beam angle and beam number error. The success rate of the revised CBR system with 10 clusters was improved by 30% with respect to the beam number error and by 13.79% with respect to beam angles error compared to the original CBR system. With the retrieved better treatment plan the beam angle adaptation should be easier to conduct.

The treatment planning machine settings have been changed in Nottingham City Hospital, which means that the beam angles in the treatment plans in the case base and the treatment plans of the six new cases obtained from the City Hospital are not directly comparable. In the future we will calibrate these cases and include them in the case base. In addition, the improved CBR system will be presented to medical physicists for further evaluation. Medical physicists adapt treatment plans by placing beams on the patient case image and adjusting them in a trial and error fashion. Our future work will investigate automated ways of treatment plan adaptation in which cases will contain expertise of medical physicists in adaptation of treatment plans. Also, the CBR system needs to be validated, therefore, we will also investigate the possibility of automated validation of the system. Finally, the system will be evaluated by medical physicists.

Acknowledgments

Authors of this paper would like to thank Angela McKenna and Louise Newton and other medical staff at Nottingham City Hospital for providing us with data, their expertise in radiotherapy treatment planning and their valuable time.

References

[1] National Cancer Intelligence Unit 2013. Office for National Statistics, UK (Accessed 02 Nov 2014).
[2] Oldham M, Khoo VS, Rowbottom CG, Bedford JL and Webb S. 1998 A case study comparing the relative benefit of optimizing beam weights, wedge angles, beam orientations and tomotherapy in stereotactic radiotherapy of the brain Physics in medicine and biology 43 21
[3] Shepard D, Ferris M, Olivera G and Mackie T. 1999 Optimizing the delivery of radiation therapy to cancer patients Siam Review, JSTOR 41 721-744
[4] Holder A and Salter B 2004 A tutorial on radiation oncology and optimization Emerging Methodologies and Applications in Operations Research (Kluwer Academic Press: Boston)
[5] Rocha H, Dias J, Ferreira B and Lopes M. 2012 Combinatorial optimization for an improved transition from fluence optimization to fluence delivery in IMRT treatment planning Optimization 61 969-987
[6] Kolodner, J. 1992 An introduction to case-based reasoning Artificial Intelligence Review 6 3-34
[7] Jagannathan R, Petrovic S, McKenna A and Newton L 2012 A Novel two phase retrieval mechanism for a clinical case based reasoning system for radiotherapy treatment planning Int. J. Artif. Intell. 21 46-70
[8] Jagannathan R and Petrovic S 2012 A Local Rule-Based Attribute Weighting Scheme for a Case-Based Reasoning System for Radiotherapy Treatment Planning Case-Based Reasoning Research and Development, Lecture Notes in Computer Science (Springer: Berlin Heidelberg) 7466 167-181
[9] MacQueen J 1967 Some methods for classification and analysis of multivariate observations Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. (University of California Press: Berkeley, CA) 1 281-297
[10] Hartigan, J. A. 1975 Clustering Algorithms (Wiley: New York)
[11] Magnetoencephalography data. MRI coordinate systems: The head coordinate system (Accessed 02 Nov 2014).
[12] Neuroanatomy Atlas. MRI brain coordinate system 2002 McCausland Center for Brain Imaging, University of South Carolina (Accessed 02 Nov 2014).
[13] Kaufman L and Rousseeuw P J 1990 Partitioning around medoids Finding groups in data: an introduction to cluster analysis 68-125
[14] Fraley C and Raftery A E 2002 Model-based clustering, discriminant analysis, and density estimation Journal of the American Statistical Association 97 611-631
[15] Calinski T and Harabasz J 1974 A Dendrite Method for Cluster Analysis Commun. Stat. 3 1-27