COMBINING DECISION TREES AND K-NN FOR CASE-BASED PLANNING

Sofia Benbelkacem, Baghdad Atmani and Mohamed Benamina

Computer Science Laboratory of Oran (LIO), University of Oran, Algeria
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algeria
{sofia.benbelkacem, atmani.baghdad, benamina.mohamed}@gmail.com

ABSTRACT

In everyday life, we are often faced with similar problems which we resolve with our experience. Case-based reasoning is a paradigm of problem solving based on past experience. Thus, case-based reasoning is considered as a valuable technique for the implementation of various tasks involving solving planning problem. Planning is considered as a decision support process designed to provide resources and required services to achieve specific objectives, allowing the selection of a better solution among several alternatives. However, we propose to exploit decision trees and k-NN combination to choose the most appropriate solutions. In a previous work [1], we have proposed a new planning approach guided by case-based reasoning and decision tree, called DTR, for case retrieval. In this paper, we use a classifier combination for similarity calculation in order to select the best solution to the target case. Thus, the use of the decision trees and k-NN combination allows improving the relevance of results and finding the most relevant cases.

KEYWORDS

Case-Based Reasoning, Classifier Combination, Data Mining, Case Retrieval, Decision Tree, Planning

1. INTRODUCTION

Planning is currently of great interest because it combines two major areas of Artificial Intelligence, exploration and logic. The intersection of these two areas has led to improved performance over the last twenty years [2]. The planning emergence in Artificial Intelligence led to the so-called classical planning [3]. But the classical planning has multiple drawbacks like the unrealistic assumptions that recognize the full knowledge of the environment, it is insensitive to changes in the environment, it does not deal with the possibility of failure or uncertainty in the environment or the presence of other agents or unpredictable situations, etc. To address these problems, a planner must be able to reason in the real world with the notion of time and resources, support more expressive representation of knowledge, evolve using past experience, cooperate with other planners, etc [4]. The rejection of the classical planning paradigm has resulted in new planning techniques aimed at solving the problems which can't be solved by traditional planning systems. Among these techniques, we are interested in case-based planning. Case-based planning is based on the reuse of past successful plans for the development of new plans. A plan for a set of objectives is not built piece by piece but by changing a memory map that partially or fully satisfies the objectives. So, the case-based planning provides significant time savings by avoiding trying to solve problems already treated. Then, to take advantage of past experience and optimize computing time, instead of synthesizing plans from primitive operators,

David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal - 2014
pp. 115–122, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41112

we adopt the case-based planning principle by combining the case-based reasoning and planning to implement our planning system guided by case-based reasoning.

In a previous work [1], we proposed a new case-based planning technique based on case-based reasoning (CBR) and decision tree, that we called DTR (Decision Tree for Retrieval). It is a decision support system that fits in the medical context. We based our approach on case-based reasoning and decision tree for many reasons. On the one hand, in care planning it is common to encounter patients who need follow the same treatment plan than others. On the other hand, the knowledge of doctors is not based only on rules but also on their theoretical knowledge and experience. For this, we use the case-based reasoning which is a paradigm of problem solving based on past experience [5]. The case-based reasoning will allow us to optimize time, given that in the medical field time is an important factor which must not be neglected. Another factor to consider in the medical field is that the data generated in health organizations are increasing. To manage a large amount of data we used data mining. We introduced an induction decision tree in the retrieval phase of case-based reasoning process. This step requires the use of a similarity measure between cases. We therefore used a retrieval phase guided by decision tree as a measure of similarity [1]. However, we cannot always rely on the solutions proposed by the retrieval by decision tree that could provide impertinent solutions if there is not enough examples (cases) in the case base. To overcome this drawback, we propose to use different classifiers and combine their predictions with the majority vote in order to achieve a more relevant result. The objective of this work is to improve the results quality of the proposed approach by using a classifier combination. Indeed, the main idea behind the combination of classifiers is an increase in the quality of results [6]. To perform the experiment, we evaluated the approach on real cases in the medical field, specifically for tuberculosis treatment.

The paper is organized as follows. In Section 2, we mention some works about the similarity measures used for retrieval. Then, in Section 3 we give a description of the proposed approach. Section 4 presents some experimental results, which include a combination of classifiers. Finally, Section 5 is devoted to conclusions and perspectives of this work.

2. LITERATURE REVIEW

The retrieval step of case-based reasoning process requires the use of a similarity measure. This notion of similarity between cases has been the subject of several works implementing various similarity measures. Nunez et al. [7] propose a new similarity measure for case retrieval. It takes into account the different nature of the quantitative or qualitative values of the continuous attributes depending on its relevance. Thus, different criterions of distance are applied for continuous attributes. Guo & Neagu [8] propose a similarity-based classifier combination system. The classifiers studied include voting-based k-nearest neighbours, weighted k-nearest neighbours, k-nearest neighbours model-based classifier and contextual probability-based classifier. Juarez et al. [9] propose a temporal similarity measure for heterogeneous event sequences, based on the overall uncertainty of a temporal constraint network. The temporal similarity is measured by describing a unique temporal scenario of temporal relations and calculating the uncertainty produced. Petridis et al. [10] present a system built on a similarity metric using a graphical representation of shapes for retrieval. The special feature of this system is that similarity is derived primarily from graph matching algorithms. Zhong et al. [11] propose a two-layer case retrieving method applied to emergency field. This method is based on structural and attribute similarity degrees. First, the structural similarity degrees between the historical cases and the current problem are analyzed. Second, the attribute similarity degrees between them are analyzed. At last, the synthetic similarity degrees between them are calculated. This method can avoid failing to calculate the similarities among the cases with the missing values and the similarity degrees between the historical cases. Hashemi et al. [12] propose a new measuring similarity method between work pieces using numeric and some symbolic attributes. This method is a similarity measurement system used for fixture design. It is composed of template retrieval and nearest neighbor. Kumar Jha et al. [13] propose a case-based decision support system for patients with diabetes. This system uses similarity by ontology to retrieve similar cases from the case base and generates a basic care plan.

A novel planning method, called DTR, is proposed here to improve the retrieval step of case-based reasoning. Figure 1 illustrates the general architecture of the proposed approach. It consists of several steps going from the project description to the data classification.

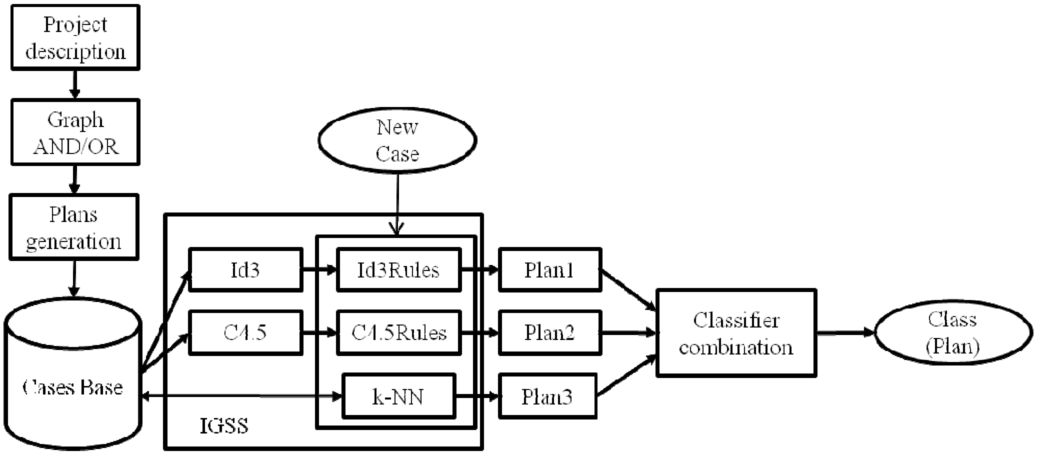

Figure 1. General architecture of the proposed approach

We call project, the set of actions to be undertaken to meet an identified need within a specified time. The organization and the sequence of tasks are generally in the form of tables or graphs. We describe the project representing the sequence of tasks in a table to generate the graph AND/OR [1]. Table 1 shows the project description of tuberculosis treatment.

Table 1. Project description of tuberculosis treatment.

Rubric	**Task**	**Description**	**Anteriority**
	Begin	Start of treatment	-
Treatments	A	Rx thoracic + bacteriologic exam	Begin
	B	2RHZE/4RH (2tablets/day)	A
	C	2RHZE/4RH (3tablets/day)	A
	D	2RHZE/4RH (4tablets/day)	A
	E	2SRHZE/1RHZE/5RHE	A
	F	3OKZE/18OEZ	A
	G	3EthOKZC/18EthOZ	A
Controls	H	Controls (I)	B, C, D
	I	Controls (II)	E
	J	Controls (III)	F, G
	End	End of treatment	H, I, J

A graph AND/OR is a graph whose nodes represent tasks and the edges represent relationships between tasks. A task represents the action performed for a period of time and relationships between tasks are constraints to satisfy [3]. The graph AND/OR given in the Figure 2 is generated from the project of tuberculosis treatment described in the previous step.

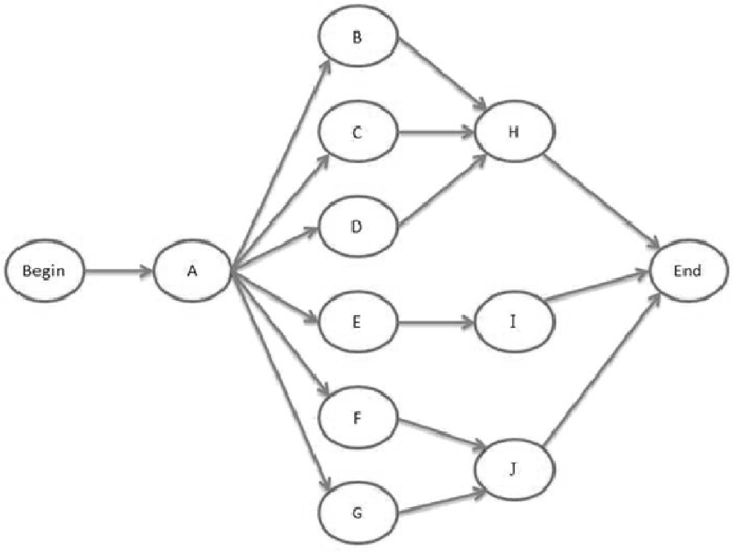

Figure 2. Graph AND/OR of project description of tuberculosis treatment

Then, planning algorithms [14] are applied to the graph AND/OR to determine the possible plans. Each path from the initial state to the final state is a plan. Let us consider the problem of planning tuberculosis treatment. In what follows, we describe the process flow of a training set (case base) consisting of cases representing patients treated for tuberculosis. Updating Id3Rules and C4.5Rules requires two samples denoted Ωa and Ωt, which are subsets of the case base. The first one, Ωa, used for training, will serve for the construction of classification rules. The second one, Ωt, used for test, will serve for testing the validity of the classification model.

These are real cases that we collected in the pneumo-phtisiology service of the University Hospital Center of Oran. Cases are described by five descriptive variables *X*1, *X*2, *X*3, *X*4, *X*5. *X*1: Sex, can take two values, *H* and *F*; *X*2: Age; *X*3: Weight, the weight of the patient; *X*4: Antecedent, can take two values, *NT* for tuberculosis and *T* if the patient has been treated for tuberculosis; *X*5: Type, can take four values, *new* if the patient has never taken treatment for tuberculosis or if he has taken for less than a month, *interruption* if the patient discontinued the treatment for two months or more, *failure* treatment failure appears for the fifth month or more, *relapse* patient declared cured but has tuberculosis again.

Each case is associated with a class *Y* which takes its values in the set of classes *C*={*T*1, *T*2, *T*3, *T*4, *T*5, *T*6} which determines the different treatments of pulmonary tuberculosis. The population ΩA considered for classification in Table 2 is a sequence of observations (or cases) *ωi* with their corresponding class *Y* (*ωi*). It's to propose an adequate treatment according to the different descriptors.

Table 2. Extract of the training set ΩA.

Ω	**X_1(ω)**	**X_2(ω)**	**X_3(ω)**	**X_4(ω)**	**X_5(ω)**	**Y (ω)**
ω_1	H	42	51	NT	New	T_2
ω_2	F	25	35	T	relapse	T_4
...	...	...	...	...	...	...

We use the data mining tool IGSS (Induction Graph Symbolic System) to build the classification model. IGSS has been developed in our research team SIF (Simulation, Intégration et Fouille de données) to enrich the graphical environment of Weka platform. It uses Boolean modeling to optimize the induction graph and automatic generation of rules [15]. Provided that the training set ΩA is representative of the original population, we can deduce classification rules which are of the form: If Condition Then Conclusion. Condition is a logical expression consisting of disjunction of a conjunction that will be called premise and Conclusion is the majority class in the node described by the condition.

The decision tree can be used in different ways [20]: classification of new data, estimation of an attribute, extraction of classification rules for the target attribute, etc. In our case, it is to classify new data. The new data is a new case which we do not know its solution part. To find the solution part we apply the retrieval step of case-based reasoning. The retrieval involves looking for similar cases to the new data. We treat this step using decision tree. The new data will be incorporated into the classification model which consists of a decision tree and classification rules. The classification model is responsible to classify new data by assigning a plan (class).

4. EXPERIMENT AND DISCUSSION OF RESULTS

During the construction of the classification model, we used only a single classifier to build the decision tree. But with one classifier and few examples in the case base, we can sometimes leads to an impertinent result. For this, we found useful to adopt a combination of classifiers in order to improve the quality of results and we test it on different datasets. First, we evaluate the proposed approach on six public datasets extracted from the UCI machine learning repository [18]. General information about these datasets is listed in Table 3. Then, we perform some experiments on a dataset of real cases which represent patients treated for tuberculosis. An extract of the case base is given in Table 2. In Table 3, the meaning of the title in each column is as follows, NA: Number of attributes, NN: Number of Nominal attributes, NO: Number of Ordinal attributes, NB: Number of Binary attributes, NI: Number of Instances and CD: Class Distribution.

Table 3. General information about UCI datasets.

Dataset	NA	NN	NO	NB	NI	CD
Glass	9	0	9	0	214	70:17:76:0:13:9:29
Hepatitis	19	6	1	12	155	32:123
Ionosphere	34	0	34	0	351	126:225
Iris	4	0	4	0	150	50:50:50
Wine	13	0	13	0	178	59:71:48
Zoo	16	16	0	0	90	37:18:3:12:4:7:9

We used the Percentage split method with a rate of 80% to evaluate the prediction accuracy of three classifiers Id3 [16], C4.5 [17], k-NN [19] and their combination. This method takes 80% of data inside the case base for the training set and 20% of the test set. K-NN is k nearest neighbors, we took k = 5 and the classifiers Id3 and C4.5 are designed for construction of the decision tree. Additionally, we adopted a combination of classifiers by majority vote. It is to count the number of votes for each class offered by different classifiers and choose the class with the highest number of votes (the most proposed class by the classifiers). We consider the C4.5 classifier as the most priority. If all classes have the same number of votes (each classifier gives a different result) then we take the proposed solution by the Id3 classifier. To assess the performance of the proposed approach, we compare our experimental results with four similarity-based classifier combination methods Maximal Similarity-based Combination (MSC), Average Similarity-based Combination (ASC), Weighted Similarity-based Combination (WSC) and MV proposed by Guo

& Neagu [8]. The experimental results are given in Table 4. Table 4 shows that the classification accuracy of the three classifiers Id3, C4.5 and K-NN is slightly better than other methods over five datasets while the other methods have a better performance with the Wine dataset. However, we note that the performance seems better with the combination of the classifiers Id3, C4.5 and k-NN than the ensemble classifiers and other methods in most of the cases.

Table 4. Comparison between classifier combination and other methods with UCI datasets.

Dataset	Id3	C4.5	k-NN (k=5)	Classifier combination	Other methods			
					MV	**MSC**	**ASC**	**WSC**
Glass	100	100	93.92	100	69.52	70.95	70.95	70.95
Hepatitis	87.5	76.12	71.61	87.5	85.33	87.33	86.67	87.33
Ionosphere	94.32	95.44	90.88	95.5	88.57	89.43	88.86	89.43
Iris	96.66	96	94.66	96.7	96.00	96.67	96.67	96.67
Wine	89.88	80.33	76.40	89.9	95.29	96.47	96.47	96.47
Zoo	98.01	99.01	95.04	99.01	95.56	95.56	96.67	96.67

Next, we calculated the metrics Precision, Recall, F-measure and Accuracy for each classifier (Id3, C4.5, k-NN) and for the combination of these classifiers applied to the real case base about tuberculosis. The performance evaluation is given in the Table 5. The experimental results presented in Table 5 show that C4.5 has a better performance than the other classifiers and the classifier combination obtains the highest Precision, Recall, F-measure and Accuracy among other classifiers on tuberculosis dataset. Thus, we can see that the classifier combination with the majority vote can improve the relevance of results.

Table 5. Performance evaluation with the case base of tuberculosis.

Evaluation metrics	Id3	C4.5	k-NN	Classifier combination
Precision	62.5	83.3	71.4	83.3
Recall	83.3	83.3	83.3	83.3
F-measure	71.4	83.3	76.9	83.3
Accuracy	80.9	90.4	85.7	90.4

5. CONCLUSIONS

We have proposed a new approach of case-based planning based on case-based reasoning and decision trees. The objective of our approach is to provide support to practitioners in selecting the appropriate treatment. We implemented our approach on real cases involving patients treated for tuberculosis. We used the IGSS tool for building decision tree and generate classification rules from the training set. To improve the quality of results, we used combination of classifiers. Thus, the use of multiple methods simultaneously can possibly afford to combine the advantages without accumulating disadvantages. For this, we combined decision trees and k-NN in order to get the mostly proposed solution (most relevant). As a method for combining classifiers, we used the majority vote because it is a fairly simple feature fusion and the most used. Faced with new data, the system classifies this data by associating a class that corresponds to a plan. To assess the performance of the proposed approach, we calculated evaluation metrics. The results of experimentation show that the performance becomes higher with the combination of classifiers. Thus, the proposed solution for the new case is more relevant. As future work, we propose to combine with other classifiers which could probably give better results. Moreover, we intend to apply this approach in another important area for further search, it is the paediatric emergency planning.

REFERENCES

[1] Benbelkacem, S., Atmani, B. & Mansoul, A. (2012) "Planification guidée par raisonnement à base de cas et datamining: Remémoration des cas par arbre de décision", Atelier aIde à la Décision à tous les Etages Aide@EGC2012, pp62-72.

[2] Bibai, J. (2010) "Segmentation et évolution pour la planification: le système Divide-And-Evolve", Doctoral Dissertation, University of Paris-sud XI Orsay.

[3] Baki, B. & Bouzid, M. (2006) "Planification et ordonnancement probabilistes sous contraintes temporelles", Actes du 15e congrès francophone de Reconnaissance des Formes et Intelligence Artificielle, pp99-107.

[4] Vlahavas, I. & Vrakas, D. (Eds.) (2005) Intelligent Techniques for Planning, Idea Group.

[5] Kolodner, J. (1993) Case based Reasoning, Morgan Kaufmann.

[6] Chitroub, S. (2004) "Combinaison de classifieurs: une approche pour l'amélioration de la classification d'images multisources/multidates de télédétection", Télédétection, Vol. 4, No. 3, pp289-301.

[7] Nunez, H., Sanchez-Marre, M., Cortes, U., Comas, J., Martinez, M., Rodriguez-Roda, I. & Poch, M. (2004) "A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations", Environmental Modelling & Software, Vol. 19, No. 9, pp809-819.

[8] Guo, G. & Neagu, D. (2005) "Similarity-based classifier combination for decision making", Proc. of IEEE International Conference on Systems, Man and Cybernetics, pp176-181.

[9] Juarez, J., Guil, F., Palma, J. & Marin, R. (2009) "Temporal similarity by measuring possibilistic uncertainty in CBR", Fuzzy Sets and Systems, Vol. 160, No. 2, pp214- 230.

[10] Petridis, M., Saeed, S. & Knight, B. (2010) "An automatic case based reasoning system using similarity measures between 3D shapes to assist in the design of metal castings", Expert Update, Vol. 10, No. 2, pp43-51.

[11] Zhong, Q., Zhang, X., Guo, S., Ye, X. & Qiu, J. (2010) "The method of case retrieving in the emergency field based on CBR", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[12] Hashemi, H., Shaharoun, A. & Izman, S. (2013) "To improve machining fixture design: A case based reasoning paradigm", Journal of Basic and Applied Scientific Research, Vol. 3, No. 5, pp931-937.

[13] Jha, M.K., Pakhira, D. & Chakraborty, B. (2013) "Diabetes detection and care applying CBR techniques", International Journal of Soft Computing and Engineering, Vol. 2, No. 6, pp132-137.

[14] Benbelkacem, S., Atmani, B. & Benamina, M. (2013) "Planification basée sur la classification par arbre de décision", Conférence Maghrébine sur les Avancées des Systèmes Décisionnels.

[15] Atmani, B. & Beldjilali, B. (2007) "Knowledge discovery in database: Induction graph and cellular automaton", Computing and Informatics Journal, Vol. 26, No. 2, pp1001-1027.

[16] Quinlan, J. (1986) "Induction of decision trees", Machine Learning, Vol. 1, pp81-106.

[17] Quinlan, J. (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

[18] Newman, D.J., Blake, C.L. & Merz, C.J. (1998) UCI repository of machine learning databases, University California, Irvine.

[19] Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. (2003) "Knn model-based approach in classification", International Conference on Ontologies Databases and Applications of Semantics, pp986-996.

[20] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) Classification And Regression Trees, Chapman and Hall, New York.

AUTHORS

Sofia BENBELKACEM is a PhD student at the University of Oran and affiliated researcher in Oran Computer Lab. Her research interests include Data mining, Planning, Case-based reasoning and Medical decision support systems.

Baghdad ATMANI received his PhD in Computer Science from the University of Oran (Algeria) in 2007. He is currently a Professor in Computer Science. His interest field is artificial intelligence and machine learning. His research is based on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic. His research are guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

Mohamed BENAMINA is a PhD student at Oran University and affiliated researcher in Oran Computer Lab. His research interests include Data mining with Ontology, Fuzzy Logic, Fuzzy Expert Systems and Fuzzy Reasoning.