Knowledge-Based Systems 24 (2011) 257–268

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Introducing attribute risk for retrieval in case-based reasoning

Juan L. Castro, Maria Navarro [*], José M. Sánchez, José M. Zurita

Department of Computer Science and Artificial Intelligence, ETSI Informática, Granada University, Spain

ARTICLE INFO

Article history:
Received 25 January 2010
Received in revised form 24 August 2010
Accepted 3 September 2010
Available online 21 September 2010

Keywords:
Case-based reasoning
Fuzzy logic
Similarity measure
Fuzzy rules
Risk

ABSTRACT

One of the major assumptions in case-based reasoning is that similar experiences can guide future reasoning, problem solving and learning. This assumption shows the importance of the method used for choosing the most suitable case, especially when dealing with the class of problems in which risk, is relevant concept to the case retrieval process. This paper argues that traditional similarity assessment methods are not sufficient to obtain the best case; an additional step with new information must be performed necessary, after applying similarity measures in the retrieval stage. When a case is recovered from the case base, one must take into account not only the specific value of the attribute but also whether the case solution is suitable for solving the problem, depending on the risk produced in the final decision. We introduce this risk, as new information through a new concept called *risk information* that is entirely different from the weight of the attributes. Our article presents this concept locally and measures it for each attribute independently.

1. Introduction

Case-based reasoning (CBR) is a well-known technique in Artificial Intelligence (AI). Attempting to imitate the way human beings reason, this technique solves problems by using or adapting solutions to old problems to solve the new ones. Although CBR has been used successfully to solve many problems and used widely, it is not entirely congruent with the way that human beings act, since the future consequences of the decisions they face constitute very important information that will be taken into account before choosing one option or another.

In real-world, resolving certain situations/problems involves some associated risk, while other situations/problems involve either no associated risk or such a small risk that it is not worth taking into account. For example, *diagnosing an illness* involves the associated risk of endangering the life of the patient, because the doctor decides the treatment of the patient, taking the diagnosis into account. For this reason, the consequences of diagnosing one illness instead of another are crucial to solving the problem. It is thus desirable to take this factor into account when facing situations in which the consequences of a decision are crucial. In contrast, other kinds of situations/problems, such as *classifying spam emails*, also involve associated risk, but the consequences of making the wrong classification are not as significant. We can thus distinguish between two main kinds of problems (Fig. 1):

(a) *Risk-problems*, where making a decision involves risk.
(b) *Non-Risk-problems*, where risk is zero or very small.

Many retrieval techniques have been developed for Non-Risk-problems. We provide a review of the research on the retrieval techniques used in CBR systems in Section 2. All types of these techniques work in a similar fashion and consist of two steps: first, the similarity between the target problem and the old cases is calculated for each attribute; second, the overall similarity is calculated as the weighted sum of similarities between attributes. Although these techniques have been used successfully to solve many problems, they are not sufficient for making the correct decision in some problems or situations (specifically in Risk situations with money or health problems, etc.). For example, when analyzing health problems, it is essential to consider whether the solution is dangerous for the patient, as we cannot risk the life of the patient. Similarly, for financial problems, it is important to consider whether the application of a solution might cause someone to lose all of his or her money. In Risk-problems, one must thus bear in mind whether the solution of the recovered case is a suitable solution to the problem in order to avoid the risk of making a significant error. In previous works [1], this risk has been taken into account by defining a global way, this approach has one problem: it cannot preserve the local information about the attribute. This article thus proposes a new local approach to solving Risk-problems, one that will find not only the most similar case but also the most appropriate solution for the case, while also improving the accuracy of the problem. We will do this by introducing new information about the problem, which we call *Risk Information*.

[*] Corresponding author. Tel.: +34 958244019; fax: +34 958243317.
E-mail address: marianj@decsai.ugr.es (M. Navarro).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.09.002

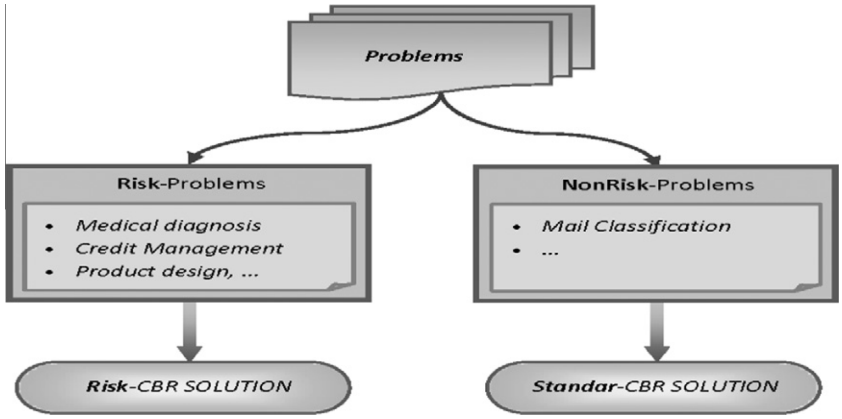

Fig. 1. Types of problems.

After applying the measures of similarity, we add a new step that uses the risk information. We call this step adequation.

Let us illustrate this idea with an example. We are interested in investing in successfully growing companies, as they seem to represent an interesting business proposition. Since case-based reasoning is useful for various types of problems and domains (as seen in [2–9]), it may be able to determine whether or not investing in a certain company will be profitable. Let us consider the following situation: Three cases, *Company A*, *Company B* and *Company C*, are stored in a case base. We know their categories and some of their characteristics. A new case, *Company D*, is entered, and we want to determine its category. The human expert has told us, however, that the attributes of *Company D* reveal that this company is not suitable for investment. Table 1 illustrates the cases, their characteristics or attributes, and the solutions.

We calculate the similarity between the cases in memory and the current case, *Company D*, to determine the category to which *Company D* belongs. This is done in two steps. In the first step, we calculate the local similarity between attributes by choosing from several measurements ([10–15]). We use the following local similarity measure:

$$sim\left(x_i^{Mem}, x_i^{New}\right) = 1 - \frac{\left|x_i^{Mem} - x_i^{New}\right|}{x_i^{max} - x_i^{min}} \tag{1}$$

where x_i^{Mem} is the *i*th attribute of the case in memory, x_i^{New} is the *i*th attribute of the current case and x_i^{max}, x_i^{min} are the maximum and minimum values between all the cases (including the target case) for the *i*th attribute, respectively.

In the second step, we calculate the overall similarity using the following arithmetic average equation:

$$Sim\left(C^{Mem}, C^{New}\right) = \frac{\sum_{i=1}^{n} sim\left(x_i^{Mem}, x_i^{New}\right)}{n} \tag{2}$$

where C^{Mem} is the case in memory, C^{New} is the target case, and *n* is the number of attributes in each case. In order to avoid confusion, the overall similarity is referred to as *Sim* and the local similarity is referred to as *sim*.

By applying Eqs. (1) and (2), we obtain the results shown in Table 2. As the table shows, the most similar case to the case under study is *Company C*. Applying the solution of this recovered case means investing in this company. We know, however, that *Company D* is unsuitable for investment (see above). Investing is thus not the best solution for the recovered case, since it does not take into account other factors that are important in choosing a case. The next step is to make the measure more accurate. We do this by incorporating the relative importance of the attributes, since certain attributes are more important than others in our problem. For example, the attribute *Cash-flow over last 5 years* does not have the same importance as the attribute *Number of employees*. We thus introduce the importance of the attributes as a new variable, which we call the weight variable. This variable measures the importance of the *i*th attribute, which we express as ω_i. Although the valuation of weights is a crucial element in determining the most similar case, our example used a human expert to assign the values. Table 3 shows the weights of the attributes.

Using the weights, ω_i, associated with each attribute in Table 3, we modify the overall similarity measure. We will thus determine the weighted sum of the similarities between attributes and weights as:

$$Sim\left(C^{Mem}, C^{New}\right) = \frac{\sum_{i=1}^{n} \omega_i \cdot sim\left(x_i^{Mem}, x_i^{New}\right)}{\sum_{i=1}^{n} \omega_i} \tag{3}$$

We calculate the similarity between the cases in memory and the new case (*Company D*) with Eq. (3) and obtain the following result: *Sim*(*Company A*, *Company D*) = 0.226, *Sim*(*Company B*, *Company D*) = 0.669 and *Sim*(*Company C*, *Company D*) = 0.6671. *Company B* beats *Company A* and *Company C* in overall similarity. As we know this to be the correct solution, our method seems to have found the solution to our problem. Without changing the order of importance of the weights, we see that small variations in the interpretation of

Table 1
Case base.

Attributes	*Company A*	*Company B*	*Company C*	*Company D*
Number of years	80	1.5	30	10
Sector	Bank	Telecom.	Distribution	Distribution
Average profit over last 5 years	3.25%	−2.15%	1%	1.5%
Listed on Stock Exchange	Yes	No	No	No
Number of employees	40,000	150	3,500	2,500
Average Cash-flow over last 5 years	10 million	−2 million	12 million	−1.3 million
Solution	Invest	No-Invest	Invest	?