FUZZY CASE IDENTIFICATION IN CASE BASED REASONING SYSTEMS

by
Nikolaidis Savvas and Lazos C.

Abstract. The most important part of a Case-Based Reasoning system is the retrieval stage, where the system must find in a sometimes-huge case base, the best matching case or cases from which to produce the prediction for the outcome of a given situation. In this paper we propose a fuzzy logic based approach for identifying cases for the similarity measuring stage of case based reasoning systems. We combine fuzzy logic with case-based reasoning because fuzzy logic is helpful for acquiring knowledge and it provides methods for applying knowledge to real-world data. A fuzzification process is implemented in a system called F-CIR and tested.

Keywords: Case-Based Reasoning, Case Retrieval, Similarity Measurement, Fuzzy sets.

1. INTRODUCTION

There are four key issues in the case-based reasoning process: (a) identifying key features, (b) retrieving similar cases in the case base, (c) measuring case similarity to select the best match, and (d) modifying the existing solution to fit the new problem. The most important part of a case-based reasoning system, is the retrieval stage, where the system must find, in a sometimes huge case base, the best matching case or cases from which to produce the prediction for the outcome of a given situation. The efficiency of this stage is a critical factor for the overall system performance. Improving retrieval is an open problem in case based reasoning research and case-based reasoning system development (Leake, 1996).

The retrieval stage requires the use of some kind of similarity measurement for the best case to match. A number of similarity measuring techniques have been used in different systems. The selection of the similarity measurement is very important, because, if the one selected is not the appropriate, the system will produce erroneous results. The Question of defining similarity is one of the most subtle and critical issues raised by case-based reasoning (Luger and Stubblefield, 1998). Very serious consideration must be given to the nature of the data, which dictate the selection of the suitable similarity measurement. The selection depends on being able to identify relevant attributes and make use of them. There is no similarity measurement that can fit all situations.

Search for similarity, is a problem which occurs in diverse applications, such as stock market prediction (Rafiei 1999, and Xia 1997), plagiarism detection (Shivakumar and Garcia-Molina 1995), forest fire prediction (Rougegrez 1993), and protein and DNA sequencing (Pearson and Lipman 1988). A number of similarity measuring techniques have been used in different systems. The selection of the similarity measurement is very important, because, if the.one selected is not the appropriate, the system will produce erroneous results. The selection depends on being able to identify relevant attributes and make use of them. There is no similarity measurement that can fit all situations. Main et al (1996) explain how fuzzy logic applies to CBR. One of the main tasks involved in the design of CBR systems is determining the features that make up a case and finding a way to index these cases in a case-base for efficient and correct retrieval. Common types of variables used to describe features in case-based systems are: Boolean, continuous, and multi-valued (ordinal, nominal, and interval-specific). Fuzzy variables allow one to represent features in an-other way: A large number of features that characterize cases frequently consist of linguistic variables which are best represented using fuzzy vectors. After testing fuzzy features in case selection, they found that the cases retrieved matched the current case in at least 95% of the tests.

Case Based Reasoning is a technique used in situations where we want to reduce the burden of knowledge acquisition, avoid repeating mistakes made in the past, work in domains where a well understood model doesn’t exist, learn from past experiences, reason with incomplete or imprecise data, provide means of explanation and reflect human reasoning (Main et al. 2000). In this paper we propose a fuzzy logic based approach for identifying cases for the similarity measuring stage of case based reasoning systems.

2. FUZZY SETS

Experts describe similarity fluently using a fuzzy vocabulary. For example, an expert may quote, “Two features are slightly similar when the difference between their values is near 10”. This kind of ambiguous knowledge is difficult to encode with classic techniques. Fuzzy sets on the other hand can do exactly that. Lotfi A. Zadeh, the founder of fuzzy logic, argues that fuzzy logic lets people *compute with word*s. He says that the fuzzy approach is necessary when the available information is too imprecise to justify the use of numbers, and second, when there is a tolerance for imprecision which can be exploited to achieve tractability, robustness, low solution cost, and better rapport with reality (Zadeh, 1996).

Fuzzy logic uses fuzzy sets to represent properties. According to classic logic every proposition must either be True or False, A or not A, either this or not this. For example, a typical rose is either red *or* not red. It cannot be red *and* not red. Every statement or sentence is true or false or has the truth value 1 or 0. Having such a property makes an item belong to a classic (also called “crisp”) set. Lotfi A. Zadeh, a professor of UC Berkeley in California, observed that conventional computer logic was incapable of manipulating data representing subjective or vague human ideas such as "an attractive person" or "pretty hot". In 1965, Zadeh published his seminal work "Fuzzy Sets" (Zadeh, 1965) which described the mathematics of fuzzy set theory, and by extension fuzzy logic. Fuzzy logic was designed to allow computers to determine the distinctions among data with shades of gray, similar to the process of human reasoning. This theory proposed making the membership function (or the values False and True) operate over the range of real numbers [0.0, 1.0]. Elements In fuzzy sets have different degrees of membership in the range from 0 to 1. Zero (0) means absolute exclusion from the set and one (1) means that the element definitely belongs to the set. All the numbers in between, declare a different degree of membership.

Let’s see an example: when can we say that the weather is “hot”? With the use of a crisp set definition we must set an arbitrary limit, let’s say 32°C. The membership function for belonging to the crisp set of “hot” is:

$$m_{hot}(x) = \begin{cases} 0 & if \quad x < 32 \\ 1 & if \quad x \geq 32 \end{cases}$$

and this function is represented in **Figure** 1:

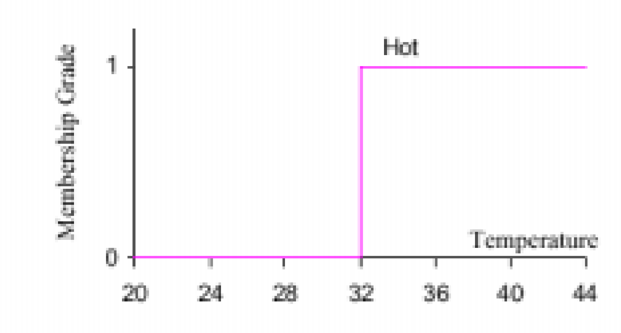


Fig. 1. Membership function for the crisp set describing “hot” weather

By using fuzzy set membership functions we can say that below 25°C the weather is definitely “not hot”, beyond 37°C it is definitely “hot” and in between the level of “hotness” increases with the temperature. Thus the fuzzy membership function is:

$$\mu_{hot}(x) = \begin{cases} 0 & if \quad x \leq 25 \\ \dfrac{x-25}{12} & if \quad 25 < x < 37 \\ 1 & if \quad x \geq 37 \end{cases}$$

and it’s graphic representation is shown in **Figure** 2.

Membership functions usually don’t have a shape as simple as “hot”. They at least tend to be triangles pointing up, as shown in **Figure** 3. This is the fuzzy set membership function for membership in the set of “warm” temperature.

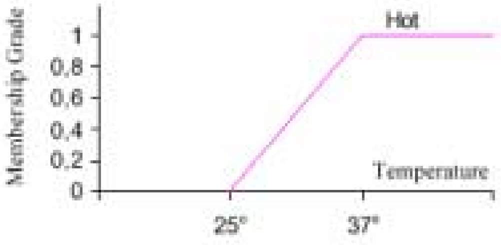


Fig. 2. Membership function for the fuzzy set describing "hot" weather

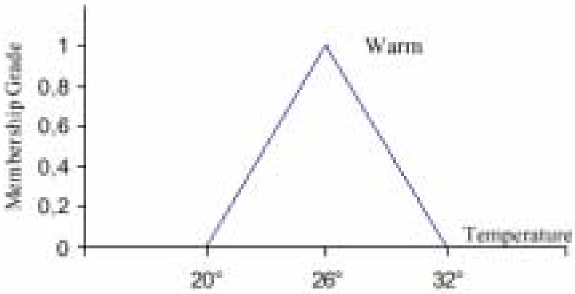


Fig. 3. Fuzzy membership function for the set of "warm" weather

Crisp sets are a special case of fuzzy sets. Most operations defined on crisp sets can also be applied to fuzzy sets. A detailed description of fuzzy logic methods is given by Zimmerman (1991).

3. FUZZY SETS COMBINED WITH CBR

Fuzzy logic is especially useful for CBR because CBR is fundamentally analogical reasoning (Leake 1996), analogical reasoning can operate with linguistic expressions, and fuzzy logic is designed to operate with linguistic expressions. We combine fuzzy logic with CBR because fuzzy logic is helpful for acquiring knowledge and it provides methods for applying knowledge to real world data. Fuzzy logic simplifies elicitation of knowledge from domain experts, such as knowledge of how similarity between two cases depends on the difference between their individual, collective, and temporal attributes. Fuzzy logic emulates human reasoning about similarity of real world cases, which are fuzzy, that is, continuous and not discrete. (Hansen, 2000).

There are at least four advantages of using fuzzy techniques in the retrieval stage (Jeng and Liang, 1995). First, it allows numerical features to be converted into fuzzy terms to simplify comparison. For example we can convert the age of a patient into a categorical scale (e.g., old, middle-aged, or young). Second, fuzzy sets allow multiple indexing of a case on a single feature with different degrees of membership. This increases the flexibility of case matching. For example, a 50-year old patient may be classified as old (0.6) and middle-aged (0.5) where 0.6 and 0.5 are the degrees that the 50-year old is classified as old and middle-aged respectively. This allows the case to be viewed as a candidate when we are looking for either an old patient or a middle-aged patient. Third, fuzzy sets make it easier to transfer knowledge across domains. For instance, we have cases showing persons older than 50 years of age (i.e. old persons) will need special effort to get a good job. We can use these cases to derive a guideline that computer software older than 2 years on the market (i.e. old software) will need special effort to make a profit. The absolute age scales are different in these two domains but the fuzzy transformation provides a bridge for comparison. Finally, fuzzy sets allow term modifiers to be used to increase the flexibility in case retrieval. For example we can search very old patients from a case base containing old patients with possibilities ranging from 0.5 to 1.0. Here “very” is a modifier of “old”, which can be used to modify the membership grade of old and result in a subset of old patients (considered very old) being retrieved. This enhances the flexibility of retrieval.

4. THE F-CIR SYSTEM: FUZZY CASE IDENTIFICATION

In this paper fuzzy logic is applied in case based reasoning systems. The system produced is called F-CIR (Fuzzy Case Identifier and Retriever) The case based reasoning system must be able to identify cases ready for retrieval. If the system is not able to properly identify a suitable case this case may not be retrieved, although it might be useful. Case adaptation may be needed for the retrieval stage to work efficiently. Case attributes can be either qualitative or quantitative. Qualitative attributes have discrete nominal values.

Case base reasoners usually work with nominal categories and problems most often occur most often occur with continuous attributes. It is necessary to devise a way to translate quantitative values to nominal ones. If we try to classify cases using classic “crisp” sets we will probably encounter problems: Classic sets do not describe qualitative attributes adequately because they give the same degree of membership to all their members and the same degree of exclusion to all the non-members. Let’s discuss for example a system that is called to decide upon the appropriate medication for a hospital patient. One factor on which the choice of the appropriate medication depends is the patient’s age. If we have a “crisp” set classification procedure then we have to set arbitrary limits to say when a man is old. This kind of classification may have its benefits but it also has its drawbacks: If we say that this limit is at 60 years of age, then for our system a 61 year old man is as “old” as a 90 year old, but in reality they are very different in their degree of oldness. A man aged 61 is *“old”* but a man of 90 is *“definitely old*”. Another problem with the traditional approach is that it does not provide adequate flexibility for marginal cases. A man aged 59 years and eleven months is classifies as a “not old” when a man aged 60 is classified as “old”, even thought they practically have the same age. This is not the way humans see things and can lead to problems. If, for example, we consider medical care in a hospital for a patient near the age limit, classifying him in one category can result in treatment not suitable for him. With fuzzy set membership functions we can have gradual membership to a set. Case based reasoners using the standard set theory are risking taking the wrong decisions by making wrong assumptions. Misinterpretations of a given situation can lead to errors. Suppose we have a system which searches for people who are “young and rich” and let’s say that someone is defined as “young” if his age is less than 30 and “rich” if his wealth is more than $1.000.000. Suppose now that we have three persons whose age and wealth are (26, $50.000), (67, $10.000.000) and (30, $999.990). None of them qualifies the criteria as being *“young and rich”*. The first is young but not rich, the second rich but not young, and the third is neither. A classic case based reasoner may retrieve the first or the second person based on partial match but it will miss the third, although in reality he is the one closest to the criteria. Using fuzzy sets can overcome the above problems. A person may have different degrees of membership in sets that would be mutually exclusive with the classic set membership definition. He can be at the same time classified as young and old, with different degrees of membership.

A two stage process is applied in F-CIR. The first stage is training the system through the *fuzzification* of the cases. The second is the traversal of the fuzzyfied case base to find the best matching case. The fuzzification works both automatically and also biased by expert tuning. Expert perceptiveness is utilized to eliminate the risk of misinterpretation of the situation. The attributes can be either discriminative or continuous. The fuzzification process is used for the quantitative, continuous attributes. The fuzzyfier describes the way a continuous attribute will be turned into a classifiable one.

The fuzzification process works according the following steps:

1. Eliminate one attribute and test (try to find irrelevant attributes).

2. Eliminate all but one attributes and test (find the relative significance of different attributes).

3. Assign weights to the attributes and normalize.

4. Identify significant or irrelevant ranges in attribute values. Enhance the former and discard the later.

5. Repeat the above steps until performance “tops up”.

The selection of fuzzy sets for representing the different categories gives the system distinct advantages:

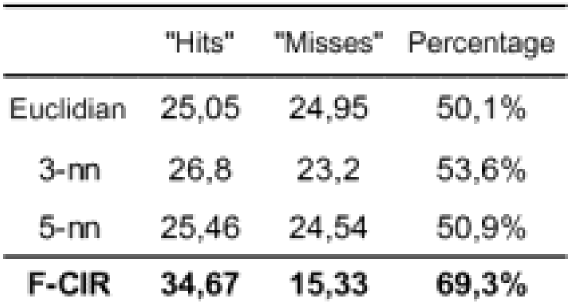
- If the attributes were categorized using crisp sets a lot of information derived from the exact value of the attribute would be lost. By using fuzzy set membership functions we preserve a lot of that information.
- Useless information that could produce erroneous predictions can be discarded. The fuzzification process can work as an expert tuning for the system.
- Having the attributes translated from quantitative to qualitative enables us to use the case base reasoner straightforwardly.
- Our prediction results become justifiable. It is easier to reason why a prediction was made based on categorical information rather than based on numbers.

5. THE EXPERIMENTS

For our experiments we used the “Boston Housing Data” from the UCI Machine Learning Data Repository. The data includes 506 cases. Each case consists of 13 parameters and a "class" attribute. The class attribute is the actual value of the house. The other attributes are:

1. Per capita crime rate by town
2. Proportion of residential land zoned for lots over 25,000 sq.ft.
3. Proportion of nonretail business acres per town
4. Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. Nitric oxides concentration (parts per 10 million)
6. Average number of rooms per dwelling
7. Proportion of owner-occupied units built prior to 1940
8. Weighted distances to five Boston employment centers
9. Index of accessibility to radial highways
10. Full-value property-tax rate per $10,000.
11. Pupil-teacher ratio by town
12. 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. % lower status of the population

Attribute 4 (Charles River dummy variable) is binary and all the other attributes are continuous, thus the fuzzification is applied to all the attributes except attribute 4. From the 506 cases of the dataset 456 are randomly selected for training and the remaining 50 are used for testing. A prediction is considered successful when the difference between the predicted and the actual value of the house is below 3000$.



	"Hits"	"Misses"	Percentage
Euclidian	25,05	24,95	50,1%
3-nn	26,8	23,2	53,6%
5-nn	25,46	24,54	50,9%
F-CIR	**34,67**	**15,33**	**69,3%**

Table I. *Mean values of correct and wrong predictions using randomly selected 50 member testing sets, according to different methods.*

The method has been tested against other standard methods such as the Euclidian distance based nearest neighbor and the k-nn technique with different values for k. After conducting a number of experiments F-CIR was correct in 69.3% of cases. The Euclidian based nearest neighbor method was correct in 50.1% of cases, 19.2% of cases less. Our system outperformed it by 38.3%. From the k-nn methods the one that performed best was for k=3. It was correct in 53.6% of cases. It classified correctly 15.7% of cases less than our method, a difference in performance of 29.3%. For other values of k the performance was even worse. For k=5, the second best performer of the k-nn methods, only 50.9% of cases were correctly classified, a deference of 18.4% from our system. F-CIR clearly outperforms the other methods. The tests have been repeated 400 times for statistical reasons and the above numbers are mean values. The results are shown in **Table I** and **Figure** 4.

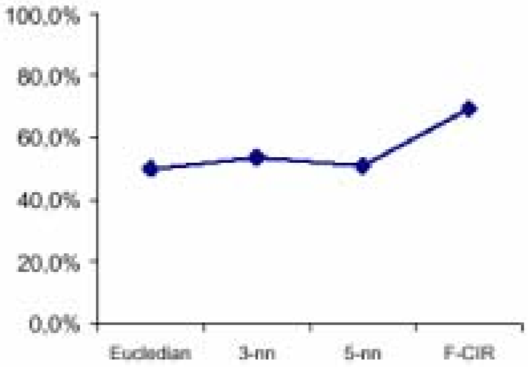


Fig. 4. Performance comparison between different methods.

6. CONCLUSIONS, BENEFITS USING FUZZY SETS

With the use of fuzzy sets, real world systems can be described naturally and accurately. Case attributes, especially attributes of a system described by an expert, are qualitative because this is the way we humans, perceive the real world. In our implementation, we give to the case based reasoning system this expert’s perceptiveness, with the use of fuzzy set membership functions for the similarity measuring stage.

Fuzzy logic enables us, to use common words as case-based reasoning attributes. Fuzzy sets can use term modifiers to modify the membership grade, define subsets and increase the flexibility of case retrieval. Terms like “very”, “somewhat” etc. can be represented and used for the similarity measurement.

By using our fuzzy logic based similarity measurement, we can increase the usability of certain types of data. Usually cases, include both qualitative and quantitative attributes that make dealing with both, more complicated. Additionally, sometimes, cases include quantitative attributes that are hard to manage. Fuzzy membership functions can be used to convert numerical attributes to quality terms simplifying the matching process and improving the system performance.

References

[1].Hansen, B. K. (2000) Weather prediction using similarity between temporal cases and fuzzy sets, Master of Computer Science thesis, Dalhousie University – Daltech.

[2].Jeng, B. C., and Liang, T.-P. (1995) Fuzzy indexing and retrieval in case-based systems, Expert Systems With Applications, Vol. 8., No. 1, 1995. Elsevier Science Ltd., 135–142.

[3].Leake, D. B. (1996) CBR in context. The present and future; in Leake, D. B. (editor) (1996) Case-Based Reasoning: Experiences, Lessons & Future Directions, American Association for Artificial Intelligence, Menlo Park California, USA, 3–30.

[4].Luger, G. F., and Stubblefield, W. A. (1998) Artificial Intelligence: Structures and Strategies for Complex Problem Solving, Addison Wesley Longman, Reading, Massachusetts, USA, pg. 238.

[5].Main. J., Dillon, T. S., and Khosla. R. (1996) Use of fuzzy feature vectors and neural ve ctors for case retrieval in case based systems, NAFIPS 1996 Biennial Conference of the North American Fuzzy Information Processing Society, IEEE, New York, NY, 438–443.

[6].Main, J.; Dillon, T. S.; and Shiu, S. C. K. 2000. A tutorial on case-based reasoning; in Pal, S. K.; Dillon, T. S.; and Yeung, D. S. eds. 2000. Soft Computing in Case Based Reasoning. London, UK, Springer.

[7].Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison, Proceedings of the National Academy of Sciences, Vol. 85, 2444–2448, April, 1988, Biochemistry.
[8].Rafiei, D. (1999) Fourier-Transform Based Techniques in Efficient Retrieval of Similar Time Sequences, Ph.D. thesis, Department of Computer Science, University of Toronto, Ontario, Canada.
[9].Rougegrez, S. (1993) Similarity evaluation between observed behaviours and the prediction of processes. In Wess, S., Al thoff, K. D. and Richter, M. (eds.), Topics in Case-Based Reasoning, Proceedings First European Workshop on Case-Based Reasoning, 1993, Springer-Verlag, Berlin, 155– 166.
[10].Shivakumar, N., and Garcia-Molina, H. (1995) SCAM: A copy detection mechanism for digital documents, Digital Libraries ‘95, The Second Annual Conference on the Theory and Practice of Digital Libraries, 155–163.
[11].Xia, B. B. (1997) Similarity Search in Time Series Data Sets, Master of Science thesis, Department of Computer Science, Simon Fraser University, BC, Canada.
[12].Zadeh, L. A. (1965) Fuzzy sets, Information and Control, Vol. 8 (June 1965), 338-353; reprinted in Bezdek, J. C., and Pal, S. K. (eds.) (1992) Fuzzy Models for Pattern Recognition, IEEE Press, 35– 45.
[13].Zadeh, L.A. – “Fuzzy Logic = Computing with Words”, IEEE Transactions on Fuzzy Systems, Vol. 4, No. 2, May 1996, pp. 103 – 111.
[14].Zimmerman, H. J. (1991) Fuzzy Set Theory and its Applications (2nd edition), Kluwer Academic Publishers.

Authors:

Nikolaidis Savvas and Lazos C., Department of Informatics, Aristotle University, Thessaloniki, Greece snikol@csd.auth.gr, clazos@csd.auth.gr