Attribute Weights Mining for Case-Intelligent System Reasoning on Similarity Rough Sets

Jianyang Li [1, 2, *], Changtong Song [2] and Qi Wang [2]

[1]School of Electrical Engineering, Zhenjiang Institute, Zhenjiang, China
[2]School of Computer Engineering, Hefei University of Technology, Hefei, China

*Corresponding author e-mail: lijianyang@sina.com

Abstract. Attribute Weights Assignment is an important method to solve real world problems full of uncertainty, for it is difficult to acquire a comprehensive formula theoretically with which so many empirical calculations have to be drilled out of domain experts. Case-Intelligent System based on CBR (case-based reasoning), which is a human creative thinking and useful reasoning model for problem-solving, can acquire prior knowledge from the former stored cases implicating decision strategy empirical and powerful, and construct a flexible system integrated with efficient machine learning methods coping with uncertainty. Attribute weights also are the key for case similarity measurement and optimal case selection in CBR cycle, so the similarity Rough Sets is proposed for case attributes reduction, knowledge obtainment and objective weights acquiring in our case-intelligent system, which performs well in real experiments on decision-making and achieves reasonable explanations.

1. Introduction

Due to our recognition limited, our world are full of uncertainty for us to explore which is a situation of our knowledge unknown for exactly describing the existing state. Uncertainty involves imperfect or unknown information and can be described in the following three aspects, (1) Decision making- a situation where the current order or nature of things is unknown, the consequences is unpredictable and credible probabilities is undesirable (2) Information theory- a certain degree to which available choices are free from constraints. (3) Statistics- a situation where neither the probability distribution of a variable nor its mode of occurrence is known.

Different experts are obliged to find out a conclusion to explain the uncertainty. for example, Vagueness is a form of uncertainty such as long/short distance where we cannot clearly distinguish the two, and this form of vagueness can be modeled by some variation on Zadeh's fuzzy logic or subjective logic [1]. Ambiguity is a form of uncertainty for word in natural language has different meaning and its interpretation depends on the writer [2,3]. Uncertainty may be a lack of knowledge of complex facts for that if we had acquired knowledge enough for those impact factors, it could have been removed with further analysis and experimentation. Due to ignorance, indolence, or the both, our learning is a process with uncertainty though we want to reach a more precise result. We must deal with the real world problems full of uncertainty, and tolerate uncertainty existing in problems to get a feasible results; moreover how to explain the results obtained from uncertain knowledge reasoning chains many researches have proposed is a more difficult task.

As the great achievement for the simulation of human analogy learning, CBR originates from human experience learning, which obtains the similar former cases and appropriately adapts to a new situation for the new problem-solving. CBR allows similar knowledge reasoning even though some attributes missing as long as they are similar in a certain degree, and knowledge space mapping is the key for analogy learning. Cases are the presentation implicit of human sense, logics and creativity, which many artificial intelligences and cognitive psychologies are exploring the synthesis-reasoning, and naturally become a common artifice when people process experiential decision-making [4-6]. So case-intelligent system based on CBR is constructed to express human learning ability for problem-solving, where attribute weights Acquirement and assignment is a effective way to conquer the complexity and uncertainty and got a reasonable results [7,8].

2. Attributes Weighting Methods

2.1. Complicated Decision-making

Many researches have done lots from different perspective, and several corresponding decision-making methods have been proposed. But they are so complicated characteristics with uncertainty-(1) uncertainty or unknown state of running environments; (2) complicated decision-making environment and development situation; (3) decision maker having the expected requirement but not a expected value; (4) undetermined candidate alternative or estimating the implementation effect difficultly; (5) establishing decision-making model difficultly. Summarized the all, uncertainty comes from the both sides, the one is that we haven't got clear knowledge for the problem, which is objective for our cognitive process; the other is that our decision-making is subjective. So case-intelligent system is chosen for the complicated decision-making from extracting and analyzing the former cases to solve similar problems [9].

CBR systems retrieve and reuse solutions from previous solved problems that have stored as cases, searching their relations by similarity measurement at a certain level, through which solves the new issues by appropriate knowledge transformation. The process of matching is based on the similar information in the case library; and the matching methods can be composed by partial similar attributes, partial matching feature, even by interpretable matching, where the chief task is the similarity measurement [10]. kNN (k Nearest Neighbor) is the foundation for case similarity measurement, let's see how the attributes weighting influence the similarity result.

Assume case $X=(X_1,X_2,...,X_n)$ with vector X_i (i=1 to n), and case $Y=(Y_1,Y_2,...,Y_n)$ with vector Y_i (i=1 to n), then their distance can be computed: $DIST(X,Y)=\sum_i D^r(X_i,Y_i)^{1/r}$. If W_i is found as the weight of attribute X_i, then $DIST(X,Y)=\sum_i W_i * D^r(X_i,Y_i)^{1/r}$. DIST(X,Y) can be measured in Hausdorff distance, Minkowsky distance, Mahalanobis distance, Manhattan distance, Euler distance (r=2), Hamming distance (r=1), etc. so the similarity between X and Y is: $SIM(X,Y)=1-DIST(X,Y)=1-\sum_i W_i * D^r(X_i,Y_i)^{1/r}$

Due to the weights changing, their similarity value may be close, and the two cases can be treated in similar state, thus it has a significant impact on the correctness of reasoning. Obviously, The attribute weights assignment in CBR system is so important that so many distance methods in different ways may determine the similarities between cases to meet the decision needs. Considering that human cognitive ability and information processing ability are finite for decision-making, attribute weighting is a powerful tool to fit for real world uncertainty naturally [11,12].

2.2. Weights Acquiring

There are many methods to determine attribute weights for different domain using, where these methods can be divided into three categories according to the original sources data: subjective weighting method, objective weighting method and combination weighting method. Subjective weighting method is a mature method to determine the weights of attributes given by domain experts, which the original data are obtained just by their subjective judgment based on experience. The commonly used subjective weighting methods include Delphi method, AHP, binomial coefficient method, minimum square method, etc. The advantage of which is that domain experts can reasonably determine the attribute weights through their long-term empirical learning according to actual decision-making problems, and the shortcoming is that evaluation results deeply relies on domain experts which each has a different weight so its application has great limitations.

Objective weighting method is proposed for its original data which should come directly from the objective environments according to the degree of relationship and influence on their attributes. The commonly used objective weighting methods include principal component analysis, entropy method, deviation and mean square deviation method, multi-objective programming method and so on. Objective weighting method has a strong mathematical theoretical basis, and mainly determines the weight according to the relationship between the original data no need with domain experts, which becomes a shortcoming especially in decision-making full of uncertainty, for the weights this method determined may be inconsistent with our subjective wishes or actual conditions. Because the most important attributes do not necessarily make the maximum difference with other attributes theoretically, the least important attribute may have large differences and then the weight determined only by the objective weighting method may have the greatest weight; and the calculation methods are mostly complicated. Moreover this method of weighting relies on actual data from the problem domain without the participation of decision-maker.

Therefore a reasonable method - combination weighting method, which combines subjective weighting method and objective weighting method in some criteria, emerges to reach the balance of empirical knowledge and original source data themselves. For it can be seen that the first has an advantage in determining the weight according to the meaning of the attribute itself without objectivity, while the latter relies on domain experts without considering the actual meaning of the attribute, integration method can reduce the subjective randomness within subjective and objective unity, and then achieve the reliable result.

2.3. Synthesis Reasoning

RBR (rule-based reasoning) is human learning ability from complicated data to induct the knowledge represented as a set of rules, like decision table which the IF (condition) THEN (action) structure specifies a relation, directive, and strategy. RBR systems such as Expert System are designed to imitate thought processes of an expert doing a particular task, and constructed using automatic rule inference, while the condition part of a rule is matched, the conclusion should be naturally drawn out. RBR generally uses specialized knowledge to solve well-defined problems, but for complex tasks the inference chains are so complicated that how a conclusion is reached and why a specific fact is needed have been bothering experts and users, and reasonable explanation for the users are difficult. Naturally the combination of RBR and CBR is expected to construct our case-intelligent system, to give their full advantages to guarantee the good performance of the system for problem-solving with the uncertain and incomplete knowledge. There are many methods to combine them in our system:

(1) parallel reasoning and choosing the better result. RBR is very suitable for regular environments, and relies on structured prior knowledge, which is the formal expression with the most efficient knowledge in intelligent systems. On the contrary, CBR system is more effective on special circumstances, individual case corrections, convenience and accumulated experience. The different focuses of the two reliable reasoning model make the two work together and coordinate with each other in parallel technology to provide better models for such parallel computing systems.

(2) Mutual conversion. transforms case knowledge and rule knowledge into one another, and uses a form of knowledge reasoning in the system. The first is that each case implies an analogy corresponding to a rule, so the case is converted into a rule, thereby forming two knowledge bases that can be used by the intelligent system, and then refining and integrating. The second is to convert the rules into cases, to

ensure that the converted case library coverage can not be less than that of the original rule base, and integrate the two case libraries through induction technology.

(3) Master-slave auxiliary reasoning. one method is dominant with the other method is auxiliary. For example, RBR is the main with CBR as the auxiliary, for which the function of the RBR system is mainly improved by CBR technology. A key question is how much coverage of the problem CBR technology can increase, and increase the functionality of the RBR system; if a similar case is found, the correction rule can also be used to find the solution; otherwise, use the rule base to infer the solution, such as deriving the solution to the problem, or return to the CBR master module, solved by human-machine dialogue or expert.

Well known as the 4R- Retrieve, Reuse, Revise, and Retain are the four main CBR processes for a new problem solving, each of them is involving in attributes weight for similarity measurement is the key to CBR system reasoning ability. CBR process as experts can also make case adoption that is much faster and more accurate with RBR assistant, for they have excellent flexibility; especially CBR can make case good explanation to users for each case is a true action.

3. Experimental results with outlook

3.1. Improved Weighting Algorithm

Rough Sets is a useful KDD tool for information processing, which can decrease the system complexity from the suitable granularity; for the traditional RS only can deal with the discrete attribute, we use Similarity Rough Sets for our Case Intelligent system to discretize the real continuous value easily and can greatly decrease the time cost for they having the same similarity measurement. According to the similarity Rough sets theory, the measurement of similarity between objects i and j in attribute a∈A is : $S_a\left(v_i,v_j\right)=1-|v_i-v_j|/|a_{\max}-a_{\min}|$, with the definition the threshold of similarity in attribute a is t(a)∈ [0,1]. Assume a attribute subset B, B⊆A, then SIMB is the similarity relation, x SIMB y while $\prod_{a\in B} W_a * S_a(a(x),a(y)) \geq t,$ t∈[0,1] and wa is the weight of attribute a; so SIMB_(X)={x∈U: SIMBx⊆ X}, $SIM_B^{-}(X)=\bigcup_{x\in X} SIM_B x$, $POS_{SIMB}(d)=\bigcup_{X\in U/d} SIM_B\_(X)$.

The importance of an attribute can be objectively computed by the rSIMB(d), which represents the classification ability of attribute subset B, for different subsets have their different classification ability. $r_{SIMB}(d)=\dfrac{card(POS_{SIMB}(d))}{card(U)}$ $K_{SIMB}(a)=r_{SIMB}(d)-r_{SIM(B-\{a\})}(d)$ $(0\leq K_{SIMB}(a)\leq r_{SIMB}(d)\leq 1)$. Source data, attribute discretization and decision table are omitted for the paper limited, from the table 1, we can compute the importance of attribute with POS(A), core(A), etc, but they are complex with time consuming, our weighting algorithm directly computes each attribute from the discernibility matrix.

Table 1. Discernibility Matrix and Attribute Grade

	X0	X2	X4	X6	X9		a	b	c	d	e	grade
X1	b,c,d,e	b,c,d	a	a,b,c,d	a,b,c	M1	1	0	1	0	1	5
X3	a,b,c,d,e	c	b,c,d	a,b	a,d,e	M2	2	3	1	2	0	4
X5	a,c,e	b,c,d,e	b,c,d,e	a,d,e	a,b,d,e	M3	4	4	5	6	5	3
X7	a,b,c,e	a,b,d,e	c,d,e	b,d,e	a,b,d,e	M4	5	8	5	7	7	2
X8	a,b,c,d,e	a,d	b,c	b,d	e	M5	2	2	2	2	2	1

- Classifying each attribute from the discernibility matrix, marked each the attribute as M1, M2… Mm, which reflects the grade m, m-1… 1, and M1 is the highest grade.
- Counting for each property item in the collection of each class, and calculating the number of occurrences of each item.

- Comparison of their counts for each attribute from M1 to Mm, if the counts are the same, the comparison of their number in the M2 must be done, and so on.
- Different importance of an attribute is emerging for their different counts in different grades, while the number of items in the count is more than any other items within each Mi, the attribute having the higher weight is computed.

From the first step {a,c,e} can be easily realized their more importance from the discernibility matrix in grade 5, the same with RS for their core attributes; but attribute a is more import than c or e, for their counts are (2,1,0) in grade 4. Although the reduction {a,b,c,e} or {a,c,d,e}, attribute b is more import than d, for their counts are (3,2) in grade 4. each attribute weight (a:0.367, b:0.078, c:0.291,d:0.063, e:0.201) can be accurately objectively, and the complexity of our algorithm also with its time consumption are much lower than the composing POS(X) and Core(X) in RS theory.

3.2. Experiments and Analysis

The data set of "Mushrooms" in our experiments is downloaded from UCI repository of machine learning databases and describes hypothetical samples corresponding to 23 species, each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. Its main information is: 8124 Instances, 22 Attributes plus the 1 decision attribute as edible (4208- 51.8%) or poisonous (3916- 48.2%), can be all simulated as binary decision, while 5244 instances are in complete attribute value and 2480 instances in missing attribute values. So the experiment are unfolded in the following three states: complete decision table, incomplete decision table and the real world decision table (that is the real decision table with attributes missing).

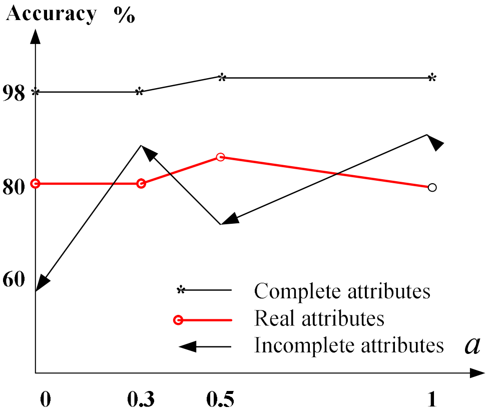

Figure 1. System Recognizing Rate with Different Weighting

We have got objective weighting method from similarity RS to decrease the influence of experts and possible errors in subjective weighting, the combines weights of subjective weighting method and objective weighting method are used in the following experiments with $I = aQ + (1-a)P$. The 10-fold cross validation is used in our experiments as usual corresponding with a=(0, 0.3, 0.5, 1).

From fig 1, we can see the first curve has little mutation, for they have the complete information on each attribute; and the pulse curve is changed sharply with a for they have the so many attributes missing for us to distinguish them and incomplete information cannot be well dealt with. As well known, fully complete or incomplete attributes are the two extreme conditions for our world full of uncertainty, combination weighting can performs well for it benefits from objective weighting and subjective weighting.

Although combination weighting can approach excellent results, it should be carefully chosen for they changes irregularly and can't get a global rule for the weight of each attribute should be determined according to the difference of each scheme under the features. Case-intelligent system ability depends on its similarity computation, simulating the reasoning and learning of human ability, though the most optimal case isn't chosen as the target sometimes, what they are mostly in candidate sets suggests us that problem-solving process should interact with us to increase system ability. The conversational CBR can be a good selection for the process of problem-solving especially in uncertain or incomplete conditions, which may assistant us in real decision- making reliably.

4. Conclusion

Attributes Weighting has been attracted much greater concerns for decision-making full of uncertainty, and how to effectively solve with such complicated or inconsistent constrains becomes a great challenge. Synthesis-reasoning technology is proposed as a solution while the domain expertise is rich with rule knowledge deficient to construct a flexible system. Combination weighting based on similarity Rough Sets can be efficiently integrated in our case- intelligent system as machine learning components, and acquires attributes preferences from the former stored cases to reduce unnecessary features and redundant information. The experiments indicate that combination weighting is a powerful method to conquer the scale and uncertainty of real problem, greatly decrease the complexity of our intelligent system and time consumption for they having the consistent similarity measurement.

Acknowledgments

This research is financially supported by the China Postdoctoral Fund of HFUT, and the Natural Science Project of Jiangsu Province YJ2017012. All the persons involved in the research projects are thanked for their help.

References

[1] Hu, Q.H., Yu, D.R., Guo, M.Z.: Fuzzy Preference Based on Rough Sets. Inf. Sci. (2011)180, 2003-2022

[2] Carmona M A, etal. Applying case based reasoning for prioritizing areas of business management. Expert Systems with Applications, 2013, 40(9): 3450-3458

[3] Sun Chia-Chi. A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods. Expert System Appl, 2010, 37(12):7745-7754.

[4] Wang H C, Huang T H. An enhanced case-based reasoning model for supporting inference missing attribute and its feature weight. Journal of Internet Technology, 2012, 13(1): 45-56

[5] Jianyang Li, Zhiwei Ni,etc, Case-based Reasonor Based on Multi-Layered Feedforward Neural Network, Computer Engineering,2006,(32)7 ,188-190

[6] Liu Y H, etal. Case learning for CBR-based collision avoidance systems. Applied Intelligence, 2012, 36(2): 308-319

[7] Yan Ai-Jun, Qian Li-Min, Wang Pu. A comparative study of attribute weights assignment for case-based reasoning. Acta Automatica Sinica, 2014, 40(9): 1896-1902

[8] Smiti A, Elouedi Z. Wcoid-Dg: an approach for case base maintenance based on weighting, clustering, outliers, internal detection and dbsan-gmeans. Journal of Computer and System Sciences, 2014, 80: 27-38.

[9] Jianyang Li, Xiaoping Liu.,Personalized Recommendation System on Massive Content Processing Using Improved MFNN. Lecture Notes in Computer Science ,Volume:7529 LNCS:183-190

[10] Tadrat J, Boonjing V, Pattaraintakorn P. A new similarity measure in formal concept analysis for case-based reasoning. Expert Systems with Applications, 2012, 39(1): 967-972

[11] Ahn H, Kim K, Man I. Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system. Expert Systems, 2006, 23(5): 290-301

[12] Lin S W, Chen S C. Parameter tuning, feature selection and weight assignment of features for case-based reasoning by artificial immune system. Applied Soft Computing Journal, 2011, 11(8): 5042-5052