The Canadian Society for Bioengineering

The Canadian society for engineering in agricultural, food, environmental, and biological systems.

La Société Canadienne de Génie Agroalimentaire et de Bioingénierie

La société canadienne de génie agroalimentaire, de la bioingénierie et de l'environnement

Paper No. 08-144

A Rule-Based Intuitive Reasoning Scheme for Exploring a Large Weather Database

Yung-Chien Sun

Bioresource Engineering, McGill University

21,111 Lakeshore Rd., Ste. Anne-de-Bellevue, Québec, Canada, H9X 3V9

yungchien.sun@mail.mcgill.ca

Robert Kok

Bioresource Engineering, McGill University

21,111 Lakeshore Rd., Ste. Anne-de-Bellevue, Québec, Canada, H9X 3V9

robert.kok@mcgill.ca

Written for presentation at the
CSBE/SCGAB 2008 Annual Conference
North Vancouver, British Columbia
July 13 - 16, 2008

Abstract

Human intuition is an inherent mental ability that has the capacity to deal with complex situations when the knowledge available is fuzzy, uncertainty, and often inconsistent. The overall objective of this work is to construct an inference engine that reasons about complex problems in a basic intuitive manner. A large weather database consisting of 54 variables and 50 years of hourly records was acquired to represent the complex situation. This database was found suitable due to its high dimensionality and the non-linear relationships among the variables. A rule-based reasoning scheme was used, with rules being derived by drawing numerous local, low certainty conclusions from small chunks of data rather than obtaining relatively fewer, global, high-certainty rules by analyzing the whole database at once. The task of the reasoning engine is to predict certain weather events by examining the body of low-quality rules and integrating corroborating evidence from them to obtain high-certainty conclusions. Two types of reasoning were tested: fast and broad. The fast reasoning mimics split-second insights and the broad reasoning represents deeper background reflection. The reasoning results are compared to those of a typical environmental data analysis approach.

Keywords: Artificial Intelligence, Intuition, Knowledge Acquisition, Weather Database

A Rule-Based Intuitive Reasoning Scheme for Exploring a Large Weather Database

Y.-C. Sun[1] and R. Kok[2]

Bioresource Engineering, McGill University, 21,111 Lakeshore Rd., Ste. Anne-de-Bellevue, Québec, Canada, H9X 3V9, Canada. [1]Email: yungchien.sun@mail.mcgill.ca, [2]Email: robert.kok@mcgill.ca

ABSTRACT

Human intuition is an inherent mental ability that has the capacity to deal with complex situations when the knowledge available is fuzzy, uncertainty, and often inconsistent. The overall objective of this work is to construct an inference engine that reasons about complex problems in a basic intuitive manner. A large weather database consisting of 54 variables and 50 years of hourly records was acquired to represent the complex situation. This database was found suitable due to its high dimensionality and the non-linear relationships among the variables. A rule-based reasoning scheme was used, with rules being derived by drawing numerous local, low certainty conclusions from small chunks of data rather than obtaining relatively fewer, global, high-certainty rules by analyzing the whole database at once. The task of the reasoning engine is to predict certain weather events by examining the body of low-quality rules and integrating corroborating evidence from them to obtain high-certainty conclusions. Two types of reasoning were tested: fast and broad. The fast reasoning mimics split-second insights and the broad reasoning represents deeper background reflection. The reasoning results are compared to those of a typical environmental data analysis approach. **Keywords:** Artificial Intelligence, Intuition, Knowledge Acquisition, Weather Database

INTRODUCTION

Intuition is, through observations, an inherent mental capability for humans, and probably for other animals with some intelligence. It plays a role in many people's day-to-day decision-making, although not always explainable and at times unsuccessful. Human intuition is known to be able to handle complex situations when and where exact, certain knowledge and other analytical methods are wanting. Intuitive reasoning, argued by Sun and Kok (Sun and Kok 2006), is a particular type of information processing, whereby low-quality evidence and causations induce a network of numerous reasoning paths and the results are obtained through integrating corroborating intermediate conclusions. From this perspective, intuitive reasoning is deemed rational and deductive, although the reasoning process usually involves little deliberation and the resulting intuitive opinions are not always explainable (Hogarth 2001).

The objective of this project is to investigate approaches for constructing a reasoning engine that can solve novel or complex problems in an intuitive manner. More details on a conceptual model of intuitive reasoning can be found in Sun and Kok (2006). The reasoning network was conceptualized as a forward-chained rule-based mechanism and the implementation was confined to one clearly defined task. Artificially generating a large set ($\sim 10^5$) of low-certainty rules from an existing database appeared a better approach than eliciting rules from human experts. To provide a sufficiently complex situation, this database was expected to be sizable, in terms of both the number of records and the attributes in each, and exhibit complicated relationships among the attributes. Given these preferences, a large weather data archive was found suitable and was acquired for this study; the task was to predict the state of certain weather variables.

Whether a problem is novel or complex is strongly dependent on a person's previous experience and existing knowledge; for a given task, if the related experiences are scant and knowledge is incomplete, that task appears to be novel or complex and an intuitive rather than deliberate inference process tend to occur. In this study, this idea serves as the basis for establishing a complex task for the proposed intuitive reasoning scheme. In real world, a problem solver rarely has access to the complete knowledge about a novel or complex task: the intuitive reasoning engine would have only partial access to the data carried in the weather archive, which was assumed to represent the body of facts necessary to the given task. Subsequently, knowledge was learned from sampling the weather database where each time a limited number of records of part of the variables were collected. Rules then were induced from individual sets of sampled data, or 'data chunks.' Because these data chunks each only reflected a portion of the whole dataset with small populations, the derived rules tended to be limited in certainty and at times conflicting.

The weather database contained tables where each row held measurements of weather variables observed at the same hour. The resulting rules described, hence, the intra-variable relationships that were rather descriptive than predictive. However, the word 'predict' is still used in this paper, as in 'the intuitive reasoning engine predicts the state of the target variable to be low,' only it means estimate instead of forecast for a future point in time. Rules induced directly from data chunks were called basic rules. Similar basic rules were bundled into super rules that represented repetitive experiences and were usually entitled to stronger confidence than average basic rules. The final rule base contained the super rules and those basic rules which were not represented by any of the former. A reasoning session where only the super rules are referenced obviously will finish earlier than one involving all the basic and super rules. These two types of inference are termed fast and broad reasoning, resembling, to some degree, the fast and lengthier information processing in reaching intuitive opinions. They were both tested and the results compared. In addition, a typical weather data analysis approach that had been applied on similar tasks was adopted. It resulted in one model for each selected target variable; the resulting models will be referred as 'typical models' hereafter. The typical models serve as the baseline of how well the relationships among the weather variables can be simulated.

THE WEATHER DATABASE

Data sets from the weather archive of Environment Canada were acquired and were assembled into a database for the years 1953-2005 for 4 Canadian sites: Montreal, St-Hubert, Vancouver, and Winnipeg. This database contains hourly observations of 54 weather variables, of which 3 variable types were identified: continuous, ordinal, and categorical. Examples for the continuous type weather variables are global solar radiation, wind speed, and temperature, for the ordinal type, levels of rain and snow, and for the categorical type, cloud type at various cloud layers. The ordinal type variables take discrete integers, the bigger the measurement the greater the intensity; zero represents 'no event.' Categorical variables also take discrete integers yet the ordering in the number does not reflect any rank of the measured quantity. Six ordinal variables were chosen (Table 1) as the target variables based on the completeness of the corresponding data in the data set.

Like other ordinary record archives, this weather data set contained missing entries. A step of imputation of missing data was applied where the missing entries were estimated by substituting values of the nearest-neighbor (Latini and Passerini 2004). Theoretically, this imputation step should not add more information to the dataset. It merely filled in empty entries with values that best fit the existing inter-variable relationships based on the same-hour measurements. Another major adjustment was done on height-related variables, such as ceiling and cloud layer heights, where number 888, amongst other measured values, flags infinite height or a cloudless sky. To avoid potential error caused by computing with infinite values, the values of these variables were converted to their reverses and all 888 entries were replaced by zero.

RULE-BASED INTUITIVE REASONING

A rule-based scheme was chosen to implement the intuitive reasoning engine because rule have been wildly adopted for knowledge representation and can be concise and easy to express (Hogeveen et al. 1994). As stated earlier, rules involved in intuitive reasoning can be characterized as low in certainty, fuzzy and inconsistent. Each of these quantities plays a certain role in developing rules in this project

Low certainty, fuzziness, and multi-valued variables

The level of certainty in intuitive reasoning is measured by a metric called strength of belief, which takes a real number in [0, 1], 0 meaning being not certain at all and 1 meaning total confidence. The strength of belief indicates the level of confidence toward the verity of a piece of information or knowledge. In other words, it applies on inputs, rules, as well as the reasoning results. The strength of belief builds up in light of corroborating evidence, e.g., pieces of evidence pointing to the same value of a given variable. The calculation for the reinforced strength of belief adopted the algebraic sum operator as in Eq. 1.

$$\begin{aligned} SB_{new} &= Reinforce(SB_1, SB_2) \\ &= SB_1 + SB_2 * (1 - SB_1) \end{aligned} \tag{1}$$

where SB_1 and SB_2 are the two original strength of belief measures that result in a higher level, SB_{new}. *Reinforce* is the reinforcement function and the algebraic sum operator ensures no strength of belief goes beyond the upper bound, 1. During rule induction, values of strength of belief were generated and assigned to basic rules when they were induced from data blocks and reinforced strength of belief calculated for super rules. During a reasoning session, the reinforcement function was also called to consolidate intermediate results. Details on the computation of the strength of belief are presented in the following subsection.

Fuzzy linguistic terms or fuzzy labels are usually used in human thoughts to specify concepts with vague boundaries. Intuitive reasoning allows the inputs as well as the conditions and the conclusion of rules to be fuzzy. The values of the ordinal and categorical variables in the database were recorded with discrete values that were taken directly as fuzzy labels representing different levels of intensity for the corresponding weather conditions. As for numerical variables, the values were converted into fuzzy labels and their associated membership degrees based on pre-defined membership functions. Inconsistency exists in a rule base as multiple-valued variables and multiple-consequent rules. When a variable is described as multi-valued, it possesses more than one possible value. This happens when information is provided by different sources. Similarly, it occurs when a rule suggests multiple consequents while the same conditions are met. This is to reflect the inconsistency that is commonly present in human knowledge but often prohibited in current rule-based expert systems. Multi-consequent rules in this study were broken down into the same number of single-consequent rules for ease of computation.

Rule learning

The plan of rule induction in intuitive reasoning followed the earlier discussion on the knowledge involved in novel or complex tasks that learning from the whole weather database at once was impossible. Alternately, rules had to be learned from distributed pieces of information (data chunks) and the learned knowledge became discrete as well. A data chunk stood for one distributed body of information from which knowledge (rules) could be learned. Each data chunks contained the values of a fixed number of variables from one 24-hour period, or one day. In this implementation, this number was set at 6 and, for each data trunk, 6 out of the 54 variables were randomly selected to be included, Fig. 1.

Basic rule derivation A 2-step procedure was implemented for basic rule induction: decision tree development and production rule generation from the resulting decision trees. In the first step, the classification version of the classification and regression tree (CART) algorithm (Breiman et al. 1984) was utilized to develop decision trees for predicting a given dependent variable from records in a data chunk. Note that not all 6 variables were used to induce trees at once. Each time, in this implementation, 1 of the 6 variables was taken as the dependent variable and 3 of the rest as the independent variables, and the records of these 4 variables became the learning record set for one decision tree. This procedure repeated for all such 4-variable combinations. The number of decision trees one data chunk resulted in thus was determined by the number of variables in a data chunk and the number of independent variables allowed in tree induction. For the current setting, every selected dependent variable would give rise to 10 trees (there are 10 different 3-variable combinations out of the remaining 5 variables) and a data chunk would result in 60 different decision trees. Note the number of variables in a data chunk and the number of independent variables included in tree induction are simulation parameters and can be changed if necessary. The former signifies how many different features one might ‘remember’ from each observation; the latter implies the maximum number of independent variables in a rule one could normally handle. Another simulation parameter about tree induction worth noting is the minimum leaf node size. Here, a leaf node was allowed to contain only one record, which later on would turn into a rule representing a rare event in memory. In the second step, rules were derived from the leaf nodes of the created decision trees. The basic rules had a production-rule like structure, comprising a number of conditions, the premise, and one conclusion, the consequent. For example, “[$7 \leq$ Humidity < 10] and [$3 \leq$ Pressure < 6] $\rightarrow$ Rain 2” reads “when the relative humidity was found to lie between classes 7 (inclusive) and 10 and the atmospheric pressure measured between class 3 (inclusive) and 6, rain of class 2 was observed.”

Through a preliminary screening step, basic rules with unsatisfactory fitting quality - the classification accuracy of the associated leaf node being lower than 50% - were singled out and discarded. A basic rule's strength of belief was decided by the size of its support, the goodness of fit of that rule, and the minimum membership degree among fuzzy labels found in its supporting records. The support was defined as the number of records in the associated leaf node times a strength-of-belief-unit, SB_u, which indicated the magnitude of support a record contributed. SB_u was supposed to be small enough so that strength of belief would not easily reach the upper bound during reinforcement, given that the strength of belief only increased in the current intuitive reasoning scheme (see the next sub-section for more discussions). For now, SB_u unit was set as 0.0001. Since in the end of reasoning the strength of belief will be interpreted comparatively, the absolute value of this unit should pose little bias on evaluating the correctness of the reasoning results.

Super rule derivation A super rule was identified from a subset of basic rules and represented a principle those basic rules collectively endorsed. First, identical basic rules were found and replaced by representative rules, which would become super rules. Next, from the rule set without replication, more super rules were created through a generalization process; for that, we adopted an instance-based rule induction approach, RISE (Pedro 1996). The input to RISE is a set of data points, called instances. The system iterates through each instance, generalizing it to cover its neighbors, and registering this generalized instance as a rule if the overall classification accuracy for the dependent variable is improved. This means assigning the covered instances the same class for the dependent variable as that of the rule leads to higher classification accuracy in general. The iteration goes on until no more improvement can be achieved. The resulting set of rules represent the original data set and will be used to classify new instances in the future. As a result, a more concise representation might be achieved without compromising the over performance of classification. Rule generalization in this study was aimed to attain similar benefits and followed the same procedure except that the instances in intuitive reasoning were already rules. For rules, to generalize means to extend the acceptable range of conditions in the premise, e.g., the rule "[7 ≤ Humidity < 10] and [3 ≤ Pressure < 6] → Rain 2" might be generalized to "[3 ≤ Humidity < 10] and [3 ≤ Pressure < 7] → Rain 2."

The strength of belief of a super rule contained 3 components: the reinforced strength of belief computed as if all involved basic rules had had identical premise, the classification accuracy of this super rule, *Acc*, and the complement of a generalization factor, F_g, as

$$SB_s = Reinforce(SBb_n,\ n = 1 \sim N) * Acc * (1\text{-}F_g) \tag{2}$$

The generalization factor measured how much this super rule had extended its original acceptable range of conditions in order to cover those basic rules. It was assumed that the higher the level of generalization, the less confidence the conclusion would assure. F_g was calculated using Eq. 3:

$$Fg = \sum_{n=1}^{N} \sum_{a=1}^{A} \sqrt{dist(a,n)} * \frac{1}{A \cdot N} \tag{3}$$

where A and N represent respectively the number of conditions in the rule's premise and the number of basic rules included. Function $dist(a, n)$ returns the square of the distance between condition a of the super rule and the *nth* basic rule; the values for all combinations are summed up and the result is then divided by the number of combinations ($A \cdot N$) to obtain the generalization factor.

The final rule set comprised super rules and the stand-alone basic rules not represented by the former. For example, a basic rule could bear such set of conditions that no super rule could cover it while still maintaining the overall performance. By and large, this rule base subsumed the frequent and rare experiences in memory, or, principles and exceptions. A typical rule base contained 32,640 rules, of which 6438 super rules. On average, the strength of belief of a super rule appeared 5 times larger than a basic rule.

The reasoning mechanism

The reasoning process proceeded in a cycle of rule search, rule firing, and integrating agreeing evidence, until no rules can be fired. A rule fired if its premise was satisfied; the strength of belief of the conclusion was determined as the minimum among that of the independent variables and of the rule. The conclusion's fuzzy membership degree was derived following the Mamdani type fuzzy inference (Mamdani and Assilian 1974), i.e., the 'max-min' composition with 'min' as the implication function. After each run of rule firing, the intermediate conclusions were examined. One character of intuitive reasoning is that it might produce multiple instances of the same class of a dependent variable, each with a distinct strength of belief and a distinct fuzzy membership degree. These dissimilar pieces of information were to be integrated for the next run. This procedure of integration is called consolidation. The consolidated strength of belief was obtained from the individual values by calling the reinforcement function. To reach an integrated fuzzy membership among different outputs was equivalent to carrying out a fuzzy information aggregation. The aggregation operators chosen here were the conventional min and max, and, to allow a degree between the two extremes, a mean operator (Loskiewicz-Buczak and Uhrig 1995). The mean operator was taken as the average weighted by the original strength of belief. For the example shown in Fig. 2, where both rules predict class 3 for the dependent variable, yet with different strength of belief, 0.6 and 0.9, and fuzzy membership degree, 0.5 and 0.8; after consolidation, the strength of belief increased to 0.96 and the fuzzy membership degree takes 3 values: 0.5, 0.8, and 0.68, one from each of the stated aggregation operators.

The difference between the fast and the broad type of reasoning lied in the rule set used for obtaining predictions. The former simulated the process of generating immediate intuitive responses by referencing only the more prominent, higher in strength of belief super rules, and the latter induced a through rule set search and took in not only the super rules but the basic rules as well. In other words, the fast type created results based on principles that were well represented in the available dataset, yet the broad type also factored in those scattered, infrequent observations that represented "exceptions" to those principles. The broad type reasoning usually required longer time periods to finish and, thus, mimicked the lengthy, reflection-like information processing.

The current intuitive reasoning scheme does not model disbelief as in MYCIN (Shortliffe, E. H. and B. G. Buchanan 1975). That is, the measures of the strength of belief monotonically increase along the process of reasoning. For example, variable 099 (Fog) has 2 possible values (classes): 0 (no fog) or 1 (fog); if evidence suggests fog is not likely to be observed, instead of lowering the strength of belief of value 1, the reasoning engine raises that of value 0. This scheme is valid due to the fact that all possible states of the weather variables are modeled in the knowledge base.

TYPICAL WEATHER DATA ANALYSIS

The typical weather data analysis in this study serves as a baseline of how well the designated target variables can be modeled by other weather variables observed on the same hour. If adequate relationships among the same-hour weather variable measurements cannot be established by a typical analysis approach, the design of this implementation scheme will be of little value. For this purpose, the first part of the technique proposed by Burrows (1999), a statistical analysis technique, was adopted as the typical weather data analysis method. The reason for choosing this technique is 2-folded. First, it is aimed to study a large weather archive and develop a prediction model for a certain target variable, which is similar to our task. Secondly, the models for ordinal target variables were built using CART, the algorithm used in rule induction for intuitive reasoning. Therefore, the results of this technique should represent a suitable benchmark for evaluating the proposed intuitive reasoning engine. The typical data analysis resulted in one model for each of the 6 selected target variables; the resulting models will be referred to as 'typical models' hereafter. Main steps for creating the typical models are listed in Table 2, with particular methods used at each step noted in parentheses.

With respect to human's use of linguistic terms instead of real numbers in processing information, the values of numerical independent variables were first converted into discrete integers, representing the different levels of intensities of the weather condition the associated weather variables described. For this, the CAIM discretization algorithm (Kurgan and Cios 2004) was adopted. CAIM is a supervised procedure where the intervals (levels) over the data range of continuous independent variables are determined so that only the minimum numbers of intervals (levels) are created while the interdependencies between the independent and the target variables are maintained. The second step was removing irrelevant independent variables which resulted in smaller data sets with only significant data records to feed to the CART algorithm. For each target variables, a pair-wise dependence test was conducted between the target variable and each of the independent variable. An independent variable was deemed as irrelevant and was excluded from the data set if the null hypothesis, the 2 variables were independent, could not be rejected with an alpha level of 0.01. If the test involved a categorical variable, the chi-square test was used; otherwise, the Kendall's Tau test was used. The rest of the procedure included calling the CART algorithm for tree induction and storing the resulting decision tree as the prediction model for the given target variables.

TESTING PLAN AND RESULTS

A five-fold simulation scheme was employed: for each city, the weather data set was randomly partitioned into 5 equal-sized subsets, each subset being the testing data set for one test run and the remaining 4 subsets used for training, and such test runs were repeated 5 times. In each run, a series of testing questions were created respectively for each of the 6 target variables from records randomly chosen from the testing set. For each question, a record was randomly selected from the test set and the value of its independent variables became the input given to the intuitive reasoning engine and the typical model. The independent variables for a given target variable were those which had been used in the associated decision tree induction for the typical model. The real class of the target variable was used to judge if a correct prediction had been made. After all questions were answered, the classification accuracy was computed.

Class imbalance problem

The weather event variables, e.g., Thunderstorm, Rain, commonly exhibit an uneven class distribution. For example, the 4 classes of variable 086, Rain, are 0, 1, 2, and 3, representing levels of rain from no-rain to heavy-rain. The ratio of occurrence among them in the training set is 1958:88:6:1; the no-rain event largely outnumbers the remaining 3 classes. The consequence of learning from this type of dataset is to have a classifier favoring the majority class while ignoring the rare event (Visa and Ralescu 2005), which is called the problem of class imbalance or class skew. A test run of the intuitive reasoning engine with rules learned from this learning set uncovered that its performance deteriorated in the presence of imbalanced class. To remedy this problem, an approach which adjusted the rules' strength of belief to offset the impact from the uneven class distribution was developed, which attempted a similar effect of one common solution for this problem: re-sampling (Estabrooks et al. 2004). A naturally infrequent class (such as heavy-rain) registers fewer records in the data set and, in turn, gives rise to fewer rules which conclude this event, as illustrated in Fig. 3. Given that a rule's strength of belief, despite other affecting quantifiers, holds a positive correlation with its support, and the sum of the support for one class, in a weak sense, measures the size of samples behind this class, an assumption was made that updating the rule strength of belief would result in a similar effect as changing the class' support or re-sampling from the weather dataset. To adjust the rule's strength of belief for one dependent variable, we first calculated the summed supports for each of its classes and calculated the median of the sums. The adjusting factor of each class was obtained by dividing the median to the individual sum, which then timed the original rules' strength of belief to produce the new measures based on the rules' predicted class. For a class originally encompassing a small support, its rules' strength of belief would increase as if they had been drawn from a bigger sample, or an up-sampled training set; for the majority classes, this worked to the effect of down-sampling.

To examine the effectiveness of this approach, intuitive reasoning was tested with both the original and the adjusted rule sets. Also, to avoid the test questions being dominated by the majority classes again, test records were chosen under supervision so that the class distribution was as even as possible, limited by the data availability of the test set.

Preliminary simulation results

By the time of submission, only a preliminary simulation had finished that contained 40 questions for each of the 6 target variables for the city of Montreal. All results presented here are averages over the 40 questions. The prediction made by the typical model contains only the class which is most likely to happen. By intuitive reasoning, on the other hand, a list of all possible classes along with the associated measures of the strength of belief and the fuzzy membership degrees are generated, as exemplified in Table 3. Note that the output of intuitive reasoning is a richer description of the predicted states than just one class code. The strength of belief reflects the relative confidence in different levels (classes) the weather event might appear and the membership degree indicates to what extent that predicted state matches the fuzzy concept of each class. For instance, the content of Table 3 may read "the reasoning engine has most confidence in the *almost* no-rain event." To compare with the results of the typical model, the class with the highest strength of belief is taken as the answer of the intuitive reasoning engine.

Compared by the average classification accuracy, the fast and the broad reasoning types generally show little difference, with either the original rule base or the rule base with adjusted rule strength of belief. However, for some target variables, the broad type appears to include more classes in its reasoning process and results. For example, for variable 091 (Snow), which had 4 classes: 0, 1, 2, and 3, the fast type only reports on the more dominant classes, 0 and 1, while the broad type also successfully pointed out the occurrence of class 2. Since otherwise there is little difference, the result of fast reasoning is used as the representative result for the intuitive reasoning engine. Figure 4 shows the averaged classification accuracy attained by intuitive reasoning with the original rule set, intuitive reasoning with rules whose strength of belief has been adjusted for the class skew problem, and the typical model, for the 6 target variables. The typical model manifests the best classification capability among the 3. For intuitive reasoning, the adjustment of rules' strength of belief appears to be helpful, or at least neutral, in improving the classification capability. Given the presence of the class imbalance problem in this task, the by-class performance was also examined. Two metrics, true positive rate (TPR) and false positive rate (FPR) calculated in a 1-vs-rest fashion, are employed for evaluation (Stäger et al. 2006). Results for variable 085, 099 and 101 are presented in Table 4. The typical model is able to identify most minor classes except the tiny population of class 3 of variable 085. Intuitive reasoning with the original rule set falls on the other end as almost failing all the rare classes. The adjustment of rule strength of belief has significantly improved the classification accuracy of intuitive reasoning for variable 099 and 101 while making no difference for variable 085; although, an inspection into the raw output file reveals that, for variable 085, the resulting strength of belief for class 2 has been drawn closer to class 0 than in the case with the original rules. This suggests that more studies on this method may further lessen the impact of the class imbalance problem in this task.

DISCUSSION

A meaningful relationship among same-hour weather variable measurements can be properly modeled by a typical data analysis scheme, at least for the majority of the selected target variables. This marks the cornerstone for the following testing and comparison because otherwise these activities would be meaningless. As for those target variables whose values were not correctly predicted, a proper model may require more weather variables or need to study more dynamic relationships, e.g., temporal autocorrelation or multivariate correlation, either case is beyond the implementation scope of this project and will not be addressed.

Secondly, the typical models outperform the intuitive reasoning engine in all categories. This result, given that the latter has access to only one ninth of the knowledge available to the former (6 variables per data chunk as opposed to 54 variables in the complete dataset), is reasonable. The complete data set symbolizes "all the truth" and the results of the typical model stand for the "real answers" that are difficult to come by for problem solvers with only incomplete knowledge of the body of truth. Therefore, the classification ability of the typical model can be deemed as an upper bound for intuitive reasoning. Note that there is still room for improvement for the current intuitive reasoning engine. The strength of belief adjustment method uniformly updates rules' strength of belief by a factor derived from the median of the by-class summed supports. Obviously this standard strategy shows different extents of effectiveness for various variables. Better performance may be gained if a level other than the median is applied; it might also be helpful to define different levels for different variables.

The third point worth noting is that the fast and broad types reasoning generate nearly identical outputs.

During reasoning, the latter differs from the former in the take-in of the basic rules. Although the basic rules seem to be often shadowed by the super rules in terms of the magnitude of the strength of belief, they do at times hint at infrequent alternates not present in the fast type thinking. This agrees with people's experience when new options dawn on them in the middle of a lengthy reflection. Fourthly, the output of classification function contains more information than the typical models. It provides all the possible results with the associated strength of belief and fuzzy membership degree. With this ensemble of information, one can review not only the outcome with the highest strength of belief but also those finishing in the 2nd or 3rd place. The fuzzy membership degree further specifies the precise position of the result within a group of similar states.

Lastly, in this project, records of same-hour measurements are studied and the models and rules are developed to estimate the target variable at the same point of time as the independent variables. It should be understood that a re-arrangement of the data set can be done to create new records containing values of different time stamps, and the subsequent models and rules will bear time-shift relationships and be able to really 'predict' the target variable: e.g., with given independent variables the reasoning engine will forecasts the target variable of the next hour. This is an ideal topic for future studies.

CONCLUSION AND FUTURE WORK

The overall objective of this project was to investigate approaches to construct an artificially intuitive reasoning engine. In this paper, the implementation of a basic intuitive reasoning engine is presented. The rule induction scheme is able to induce rules from discrete chunks of data using a common knowledge discovering technique: decision tree. The rule-based intuitive reasoning engine is capable of solving questions with a set of rules which is fuzzy, inconsistent, and low in certainty, the characteristics of the knowledge usually drawn on for dealing with complex situations. It would be interesting to further implement more methods to enrich the current rule generation repertoire, e.g., multiple polynomial regression, time series analysis, or association rule exploration, and see how or if the reasoning engine would gain in performance.

Preliminary test results show that the intuitive reasoning engine fairs reasonably well for a classification task with a large weather database. Although the problem of class imbalance is found to exist and deteriorate the classification capability of the reasoning engine, the adjustment of rule strength of belief has shown to improve the performance for a number of target variables. Study on a more sophisticated adjustment algorithm might provide further improvements. A full-scale test simulation for all 4 cities is planned and a better picture of the performance of the proposed intuitive reasoning engine will be examined and reported.

REFERENCES

Breiman, L. and J. Friedman, C. J. Stone, and R. A. Olshen. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Burrows, W. R. 1999. Combining classification and regression trees and the neuro-fuzzy inference system for environmental data modeling. In *Proceedings of the Eighteenth International Conference of the North American Fuzzy Information Processing Society* 695-699, New York, NY: IEEE Inc.

Demirli K., S. X. Cheng, and P. Muthukumaran. 2003. Subtractive clustering based modeling of job sequencing with parametric search. *Fuzzy Sets and Systems* 137(2): 235–270.

Pedro S. 1996. Unifying instance-based and rule-based induction. *Machine Learning* 24:141-168.

Estabrooks A., T. Jo, and N. Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20: 18-36.

Hogarth, R. M. 2001. *Educating intuition*, Chicago, IL: The University of Chicago Press.

Hogeveen, H., M. A. Varner, D. S. Bree, D. E. Dill, E. N. Noordhuizen-Stassen, and A. Brand. 1994. Knowledge representation methods for dairy decision support systems. *Journal of Dairy Science* 77: 3704-3715.

Kurgan L. A. and K. J. Cios. CAIM discretization algorithm, 2004. *IEEE Transactions on Knowledge and Data Engineering* 16(2): 145-153.

Latini G. and G. Passerini. 2004. *Handling missing data: applications to environmental analysis*, Billerica, MA : Computational Mechanics Inc.

Loskiewicz-Buczak A. and R. E. Uhrig. 1995. Information fusion by fuzzy set operation and genetic algorithms *Simulation, 65(*1*)*, 1995, 52-66.

Mamdani E.H. and S. Assilian. 1974. An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. *International Journal of Man-Machine Studies* 7: 1-13.

Stäger M., P. Lukowicz, and G. Tröster. 2006. Dealing with Class Skew in Context Recognition. In *Proceedings of the 26th IEEE International Conference on Distributed Computing Systems Workshops* 58-63, New York, NY: IEEE Inc.

Sun Y, and R. Kok. 2006. A conceptual model of an artificially intuitive reasoning engine for ecosystem control. In *Proceedings of the Second IASTED International Conference on Computational Intelligence*, Track #: 523-050, Calgary, AB: ACTA Press.

Visa S. and A. Ralescu. 2005. Issues in mining imbalanced data sets - a review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference* 67-73, Madison, WI: Omnipress.

Table 1. Selected target variables

Variable Code	Variable Name
085	Thunderstorm
086	Rain
088	Drizzle
091	Snow
099	Fog
101	Smoke

Table 2. General flow for typical weather data analysis

For each target variable

Discretize numerical independent variables (CAIM, Kurgan and Cios 2004)

Calculate dependence between the target variable and each independent variable (Kendall's Tau test, Chi-square test)

Remove independent variables with no or low dependence with the target variable

Generate the classification tree for this target variable (CART)

End For

Table 3. A typical output of intuitive reasoning where possible classes of the target variable are presented along with their measures of the strength of belief and fuzzy membership degrees

Variable	Class	Strength of Belief	Fuzzy Membership Degree
16	0	0.72	0.92
16	1	0.63	0.90
16	2	0.32	0.70

Table 4. True positive rate (TPR) and false positive rate (FPR) comparison by variable and their classes for the typical model, intuitive reasoning with the original rule set (IR original), and intuitive reasoning with adjusted rule strength of belief (IR adjusted). The count indicates the number of testing questions for each class.

			IR original		IR adjusted		Typical model	
Varible	Class	Count	TPR	FPR	TPR	FPR	TPR	FPR
085	0	20	1	1	1	1	1	0.4
085	2	18	0	0	0	0	0.6	0.1
085	3	2	0	0	0	0	0	0
099	0	20	1	1	1	0.6	1	0
099	1	20	0	0	0.4	0	1	0
101	0	20	1	0.9	1.0	0.6	1	0.1
101	1	20	0.1	0	0.5	0.1	1.0	0

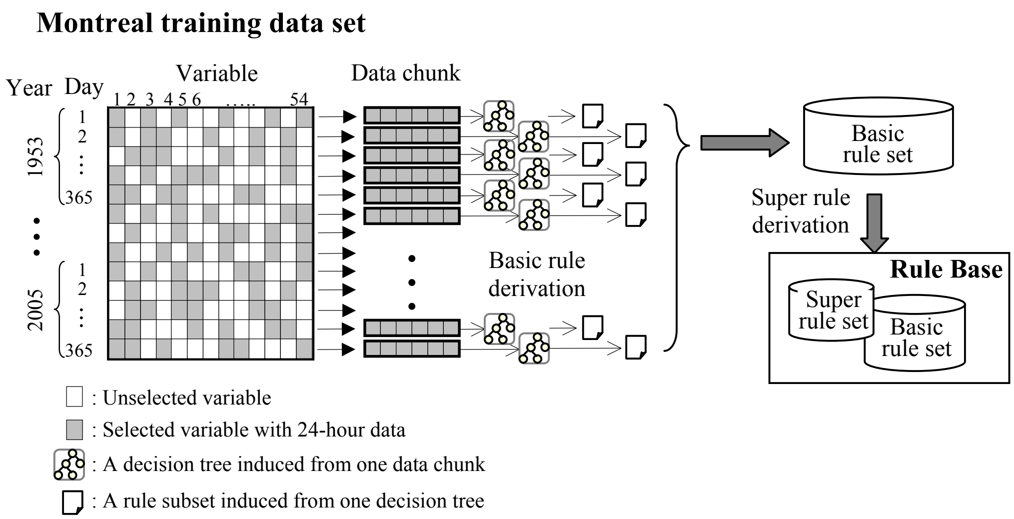

Fig. 1. Illustration of the rule learning process, where measurements of 24 consecutive hours for 6 selected variables formed a data chunk; from individual data chunks basic rule were generated through a 2-step procedure; basic rules then were gathered to become the basic rule set; after super-rule derivation, the final rule base contains super rules and those basic rules not represented by any super rules.

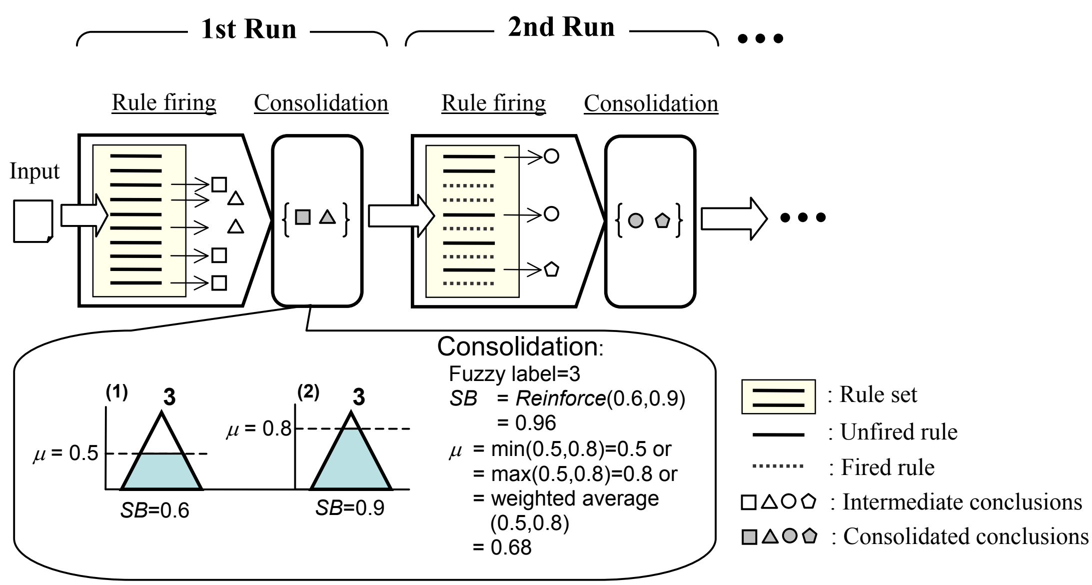

Fig. 2. Rule firing and consolidation in reasoning. Each run contains rule firing and consolidation and the reasoning process goes on until no rules can be fired. Instances of the same intermediate conclusions are consolidated; for this example, the rule firing in the first run results in 2 instances for class 3 with distinct strength of belief and fuzzy membership degree that were consolidated to “class: 3, strength of belief: 0.96, fuzzy membership degree: 0.5 (with min operator), 0.8 (with max operator), and 0.68 (with weighted average operator).”

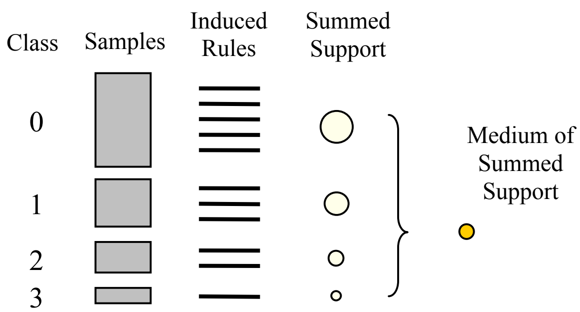

Fig. 3. The phenomenon of class imbalance in the data set and in the rule set, which illustrates that, for a variable, the uneven class distribution in the training set is carried out to the learned rule set in terms of the summed rule support by classes concluded by the rules. The poor classification performance might be improved by adjusting individual rules' strength of belief as if rule's support were adjusted so that the new summed support for all classes were brought to be equal to their original medium.

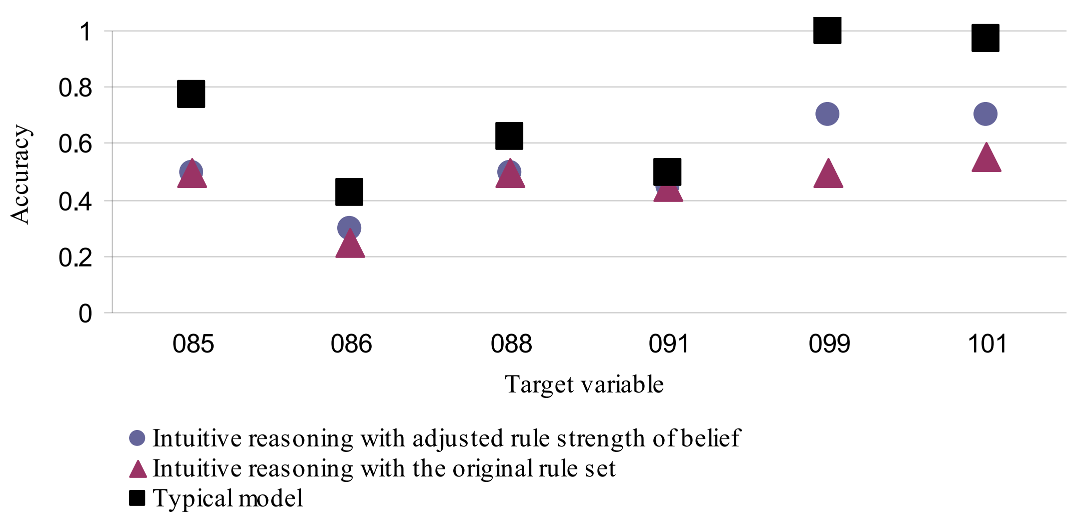

Fig. 4. Averaged classification accuracy for the 6 target variables of the typical model, intuitive reasoning with the original rule base, and intuitive reasoning with the rule base with adjusted strength of belief