Interactive Case Based Reasoning Considering Proximity from the Cut-off Point: Application to Diagnose for Diabetes

Yoon-Joo Park[a*], Byung-Chun Kim[a]

[a]KAIST, 207-43 Cheongryangri-dong, Dongdaemoon-gu, Seoul 130-722, Korea,
Tel:82-2-958-3697, Fax: 82-2-958-3604, E-mail : mirage20@kgsm.kaist.ac.kr
bckim@kgsm.kaist.ac.kr

Abstract

Case based reasoning (CBR) models often solve problems by retrieving multiple previous cases and integrating those results. However, conventional CBR makes decisions by comparing the integrated result with the cut-off point irrespective of the degree of the adjacency between them. This can cause increasing misclassification error for the target cases adjacent to the cut-off point, since the results of previous cases used to produce those results are relatively inconsistent with each other. In this article, we suggest a new interactive CBR model called Grey-Zone Case Based Reasoning (GCBR) that makes decisions focusing additional attention on the cases near the cut-off point by interactive communication with users. GCBR classifies results automatically for the cases placed outside the cut-off point boundary area. On the other hand, it communicates with users to make decision for the cases placed inside the area by verifying characteristics of the dataset. We suggest the architecture of GCBR and implement its prototype.

(**Keywords**: Interactive Case Based Reasoning; Cut-off Point; Boundary Area; Proximity; Correlation; Artificial Intelligence.)

1. Introduction

Case based reasoning (CBR) uses the results of previous similar cases to solve current problems. The results of retrieved cases can be different from each other in many circumstances, thus it classifies the results by integrating the previous results and comparing them with the cut-off point. At this stage, conventional CBR does not consider a degree of the adjacency between them. For example, let us represent the absence of disease as 0 and the presence of disease as 1, and set the cut-off point as 0.5. If there are two cases that have the integrated result 1 and 0.501, then conventional CBR classifies both of them as the presence of disease even though their levels of consistency are very different. This can increase the risk of misclassification for the target cases adjacent to the cut-off point, since the results of previous cases used to produce those results are relatively inconsistent with each other. In the previous example, the result, 1 means every previous similar case shows the presence of disease. On the other hand, the result 0.501 means the number of previous neighbors that have the presence of disease is slightly higher than the number with the absence of disease. Thus, it seems plausible to pay more attention to the cases placed near the cut-off point for classifying the results.

In this article, we suggest a new interactive CBR model called Grey-Zone Case Based Reasoning (GCBR) that can make decisions paying additional attention to the cases near the cut-off point by interactive communication with users. GCBR classifies results automatically for the cases placed outside the cut-off point boundary area. However, it communicates with users for the cases placed inside of the area by understanding characteristics of dataset. In order to do this, we introduce the concept of certainty-percentage. Certainty-percentage means the degree of consistency in terms of how much the selected neighbors produce consistent results. If every selected neighbor produces the same results than the certainty-percentage of that result becomes 100%; on the other hand, if there isn't any consistent result among the selected neighbors, then the result is situated at the exact cut-off point and the certainty- percentage becomes 0%.

The rest of this paper is organized into four chapters. Chapter 2 presents the related research. Chapter 3 introduces the new interactive case based reasoning method, called Grey-Zone Case Based Reasoning (GCBR). Next, chapter 4 shows the prototype of GCBR. Finally, a summary and concluding remark are presented in chapter 5.

* Corresponding author. Tel.: +82-2-958-3697; Fax: +82-2-958-3604; E-mail:mirage20@kgsm.kaist.ac.kr

2. Research Background

2.1 Case Based Reasoning

Case Based Reasoning (CBR) is an approach for solving a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation (Aamodt and Plaza, 1994). This concept assumes that similar problems have similar solutions, so CBR is an appropriate method for a practical domain focused on real cases rather than on rules or knowledge to solve problems. A general CBR cycle is described by the following four processes by Aamodt and Plaza (1994).

1. RETRIEVE the most similar case or cases.
2. REUSE the information and knowledge in that case to solve the problem.
3. REVISE the proposed solution.
4. RETAIN the parts of this experience likely to be useful for future problem solving.

In the retrieve process, many CBR models retrieve multiple similar neighbors rather then the single nearest neighbor. The results of the retrieved neighbors can be different from each other, thus CBR uses integrated results considering the degree of similarity and the number of neighbors. After that, it makes classification decisions by comparing the integrated results with the cut-off point. Figure 1 shows the procedure of the general CBR model.

Step 1. Begin with target case $X(t)$.

Step 2. Seek the J neighboring cases $X(t_i)$ in the past which are closest to $X(t)$ according to the distance function:

$$d_i = d[X(t_i), X(t)]$$

Step 3.Compute the sum of weights: $d_{TOT} = \sum_{i=1}^{J} d_i$

Step 4. Determine the relative weight of the i^{th} neighbor:

$$w_i = \frac{1}{J-1}\left[1 - \frac{d_i}{d_{TOT}}\right]$$

Step 5. Find the result $O(t_i)$ of each case $X(t_i)$ in the set of neighbors.

Step 6. Calculate the result $\hat{O}(t)$ of the target case X(t) as the weighted sum of results:

$$\hat{O}(t) = \sum_{i=1}^{J} w_i O(t_i)$$

Step 7. Compare the result $\hat{O}(t)$ with the cut-off point and decide classification result.

Figure 1. The procedure of general CBR using multiple neighbors

One of the issues for using conventional CBR is how many previously experienced cases to retrieve, since it can strongly influence the performance of CBR. Generally many CBR models use a fixed number of neighbors without considering the optimal number for each target case, thus it does not guarantee optimally similar neighbors for various target cases. This leads to the weakness of lowering the predictability due to deviation from desired similar neighbors. Chun and Park (2005) suggested a CBR model that adapts the optimal number of neighbors for each target case, however they focus on the number of neighbors rather than similarity. Park el al. (2006) suggested a case extraction method called Statistical CBR (SCBR) that retrieves the optimal number of neighbors based on the probabilistic similarity. The brief outline of SCBR can be summarized as follows (Park et al., 2006).

• ***The Outline of SCBR***

Step 1. Scale the data.
Step 2. Learn the distribution of distances using the learning dataset.
Step 3. Find the optimal cut-off probability from the learning dataset.
Step 4. Select the neighbors within the distance threshold calculated from the obtained optimal cut-off probability for the validation dataset from Step 2 and Step 3.
Step 5. Perform CBR using the selected neighbors and calculate the result.

2.2 Interactive Case Based Reasoning

Interactive case-based reasoning (CBR) has been a topic of interest since early 2000. Interactive CBR is an extension of the CBR paradigm in which a user is actively involved with the inferencing process (Aha et al, 2001). This paradigm was initiated due to commercial requirements rather than academic concerns, since reflecting user opinions in the CBR models can produced more customized results and gives more satisfaction to the users. However, from an academic perspective, this makes it difficult to evaluate the performance of interactive CBR models because the performance can be changed by user selection.

Although, there is a gap between academic interest and commercial concerns, the necessity of practical use of interactive CBR in the research area should not be overlooked. There is some previous research on interactive CBR. Mekenna and Smyth (2001) suggest an interactive visualization tool for case-based reasoners. This approach behind the CASECADE authoring system, which allows case authors to interact with, and be guided by a model of case competence through a variety of novel visualization tools (Mekenna and Smyth, 2001). Leake and Wilson (2001) do research on supporting aerospace design by integrating a case-based design support framework with interactive tools for capturing expert design knowledge through "concept mapping". They try to provide a useful design aid and develop general interactive techniques to facilitate case acquisition and adaptation (Leake and Wilson, 2001). In order to communicate with users, the visualization tool or interface is essential for Interactive CBR. Yang and Wu (2001) present the interactive user-interface component of the CASEADVISOR system that helps to compress a large case base into several small ones. They achieve it by merging similar cases together through a clustering algorithm (Yang and Wu, 2001). Mchserry (2001) presents a prototype environment for interactive CBR in sequential diagnosis, called CBR Strategist, which is designed to meet the user-interface requirements of intelligent systems for sequential diagnosis (Mchserry, 2001). Aha (1998) tries to improve retrieval efficiency without sacrificing retrieval precision by dialogue inferencing in conversational CBR (Aha, 1998). There is some other previous research that focuses on trying to communicate with users during CBR processes (Goker, 2000; Jurisica, 2000; Simazu, 2001; Aktas, 2004).

3. Grey-Zone Case Based Reasoning

In this chapter, we suggest a new interactive CBR method called Grey-Zone Case Based Reasoning (GCBR) that leaves the results of target cases placed in the cut-off point boundary area "undecided" to let the user make decisions individually though more analysis. We name this area a "grey-zone" since it is in the middle of definite results, white and black metaphorically, such as the presence and absence of disease or good credit and bed credit.

CBR finds the results of target cases from the previous similar neighbors. In some cases the results of neighbors are consistent but in others they are not. The concept of this paper is initiated from curiosity as to whether there is any relationship between the consistency of the results and classification accuracy. Intuitively, it seems plausible to assume that the more consistent the results of the previous neighbors are, the more accurate the classification results of the target cases are. Then, whether the average certainty-percentage of the correctly classified group is higher than those of the misclassified group is also questionable. Those questions are important, because if the assumptions are true then the target cases near the cut-off point should be treated more carefully.

In order to get the answers for these questions, we perform preliminary analysis. The results of this analysis are introduced in section 3.1. Based on this analysis we structured the architecture of GCBR. The suggested architecture of GCBR is introduced in section 3.2. Finally we explain the algorithm of overall GCBR in section 3.3.

3.1 Preliminary Analysis

There are two assumptions we try to figure out. The first assumption is that the certainty-percentages of the results are positively correlated with classification accuracies. The high certainty-percentage means many previous similar cases support the result, thus the result seems more reliable from the past viewpoint. In order to verify this assumption we perform a correlation analysis between certainty-percentage and accuracy. The second assumption is that the average certainty-percentage of a correctly classified group is higher then those of a misclassified group. The reasoning behind this assumption is deductive. In order to find out the answers for this, we divide the results into two groups and execute a t-test.

In these experiments, we use diabetes dataset taken from the UCI repository on machine learning (Blake & Merz, 1998). The dataset were collected by the National Institute of Diabetes and donated by V. Sigillito. The dataset originally contained 768 cases and 9 attributes, but we use 760 cases to construct 10 subsets of equal size. It consists of 2 classes where 492 cases show the presence of diabetes and 268 cases when it is absent. We use the first 8 attributes to diagnose diabetes and compare the results with the final attribute.

3.1.1 Correlation between certainty-percentage and accuracy

In this section we try to figure out the correlation between certainty-percentage and accuracy. In order to do this, we calculate accuracies using the cases that satisfy the certainty-percentage's upper boundary as it changes from 1% to 100%. For example, if the certainty-percentage limit is 30% then the cases that have less certainty-percentage than this limit can be used for calculating the accuracy. Figure 2 shows the graphical depiction of the accuracy of the cases that meet the limit of these certainty-percentages for diabetes dataset.

After that, we perform correlation analysis to find out if there is any relationship between certainty-percentage and accuracy. Correlation analysis measures the degree of association between two variables. We use Perarson's correlation coefficient value in this experiment, since it is the most common measure that reflects the degree of the linear relationship between two variables. It ranges from 1 to -1, which means perfectly positively correlated to perfectly negatively correlated. Table 1 shows the results of correlation analysis using diabetes datasets. The null hypothesis is that there is no correlation between certainty-percentage and accuracy. The result of this analysis indicates that there is a strong positive correlation between certainty-percentage and accuracy for diabetes dataset, since the correlation coefficient over 0.6 is usually regarded as a strong positive correlation.

	Diabetes
Pearson correlation coefficient	0.92741
P-value	***<.0001***

Table 1. Overview of correlation analysis

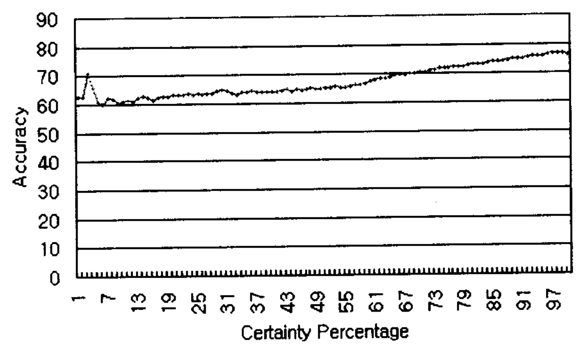

Figure 2. Accuracy of PCBR as the certainty-percentage changes (Diabetes)

3.1.2 Comparison of a correctly classified group and a misclassified group

The second assumption is the average certainty-percentage of a correctly classified group is significantly higher than a misclassified group. In order to verify this, we classify the results into the correctly classified group and misclassified group and then perform a two-sided t-test on unpaired means with unknown variance. The null hypothesis is the average certainty-percentages of the correctly classified group and misclassified group are equal. Table 2 shows the result of the t-test. It indicates that the average certainty-percentage of the correctly classified group is significantly higher than for diabetes dataset at the 95% confidence interval. We also display the frequencies of cases as they fall into each of the specified certainty-percentage categories using a histogram to provide get detailed information of the results. Figure 3 shows the histograms of the correctly classified groups and Figure 4 shows the histogram of misclassified groups for diabetes dataset using the statistical tool, eviews. The X-axis of the histograms is certainty-percentage and the Y-axis is frequency of cases. As you see in Figure 4, there are more cases adjacent to 0 certainty percentage for misclassified datasets. Conclusively, the average certainty-percentage of a correctly classified group is higher than the other and there are more cases near the cut-off point.

	Correctly classified group	Misclassified group
Obs. #	580	180
Mean	57.328	40.754
Std. Dev	30.709	30.824
T-result	***5.11E-10***	

Table 2. Overview of each group and the t-test results

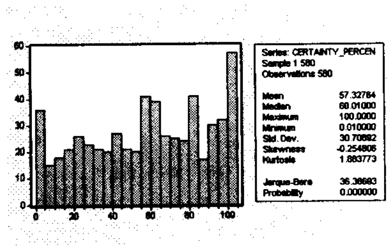

Figure 3. Histograms for diabetes (Correctly classified group)

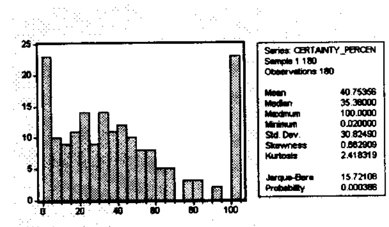

Figure 4. Histograms for diabetes (Misclassified group)

3.2 The architecture of Grey-Zone CBR

Grey-Zone Case Based Reasoning (GCBR) leaves the results of target cases near the cut-off point "undecided" to let the user make decisions individually though additional analysis. Then we have to determine the desirable grey-zone. In order to make the decision, it is necessary to consider both the user's needs and the characteristics of the datasets. The level of grey-zone means the degree of analysis that the user has to do, thus it is necessary to find out the costs that the user is willing to pay for setting the grey-zone. Also, understanding the characteristics is essential. The relationships between certainty-percentages and accuracies are different according to each dataset as we explained in section 3.1.

In this section, we explain the architecture of GCBR. GCBR is the structure for the purpose of providing effective information for users and making useful recommendations then reflecting the user's final decision in the model.

The outline of GCBR is: First, provide the information that helps user understand the characteristics of the dataset. GCBR gives the information about the previous two assumptions, that the certainty-percentage of the results is positively correlated to classification accuracy and the average certainty-percentage of the correctly classified group is higher than the misclassified group. Second, the user sets their recognized costs for each of the three cases – correctly classified, misclassified, and undecided cases. Third, GCBR suggests the grey-zone that minimizes the total cost using the following formula:

$$c_{tot} = c_{correct} \times n_{correct} + c_{wrong} \times n_{wrong} + c_{grey} \times n_{grey}$$

(c: cost of each case, n: the number in each case)

Fourth, the user set the grey-zone in reference to the information. Fifth, GCBR calculates the results and classify them except for the cases placed in the grey-zone area. The cases inside grey-zone area remain as "undecided". Sixth, GCBR supports the user in making the appropriate decision for the cases inside grey-zone. In order to support the user in this stage, GCBR gives the information on which neighbors are used for the "undecided" results and lets the user add or remove neighbors. Seventh, the user makes the final decision. Figure 5 describe the outline of SCBR and Figure 6 shows the architecture of GCBR graphically.

Step 1. Provide information for the user to understand the characteristics of the dataset.
- Give the information as to whether the certainty-percentage of the results is correlated to classification accuracy by Pearson's correlation test.
- Give the information as to whether the average certainty-percentage of the correctly classified group is higher than the misclassified group by the t-test.

Step 2. Get the recognized costs of each of the three cases: correctly classified, misclassified, and undecided cases - from the user.

Step 3. Suggest the grey-zone for the user that minimizes the total cost.

Step 4. Let the user set the grey-zone area.

Step 5. Calculate the results and classify them except for the cases placed in the grey-zone area.

Step 6. Support the user in making the appropriate decision for the cases inside of grey-zone area.
- Provide the information on which neighbors are used for the cases that are classified as "undecided".
- Give the function that inserts or removes neighbors for the "undecided" cases and recalculate the result to support the user's final decision.

Step7. Let the user make the final decisions for the "undecided" cases.

Figure 5. The Outline of GCBR

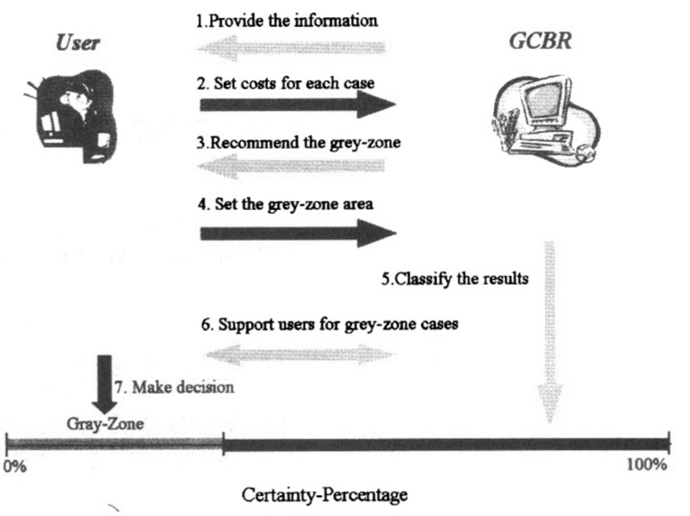

Figure 6. The architecture of GCBR

3.3 The algorithm of Grey-Zone Case Based Reasoning

The algorithm of Grey-Zone CBR (GCBR) is described in this section. In the first step, it interacts with the user to get the suitable grey-zone for the situation. In order to do this, it provides the histograms of certainty-percentage and recommends the grey-zone area. In the second step, it selects the neighboring cases in the learning dataset that satisfies the optimal cut-off probability criterion and calculates distances between neighbors and the target case. We apply SCBR to find the optimal neighboring cases in this stage. In the third step, it sums up the distances. In the fourth step, it determines the relative weight of the neighbor. In the fifth step, it sums up each weight into the same output classes. If the number of output classes is c, then the relative weight is summed up in c kinds of classes. In the sixth step, it selects the class that has the highest sum of weight and calculates the certainty-percentage. If the selected class is s, then the target case could be classified as s, unless the certainty-percentage is less then the grey-zone area. In the seventh step, it separates the cases inside the grey-zone area as "undecided" and classifies the rest of the cases. Finally, it interacts with the user and reflects the final decision of the user in GCBR. In order to do this, it provides information on which neighbors are used for the result of the target case classified as "undecided". It also provides an additional function to add or remove neighbors and recalculates the result to support the user's decision. But the user makes the final decision. The algorithm for the GCBR is presented in Figure 7.

Step 1. Interact with the user and get the grey-zone set by the user. Suggest the grey-zone that minimizes the total cost c_{tot}, however the final decision is made by the user:

$$c_{tot} = c_{correct} \times n_{correct} + c_{wrong} \times n_{wrong} + c_{grey} \times n_{grey}$$

(c: cost of each case, n: the number of each case)

Step 2. Seek the J neighboring cases $x(t_i)$ in the past which are closest to the target case $x(t)$ according to the distance function. : $d_i \equiv d[x(t_i), x(t)]$

Step 3. Compute the sum of distances. : $d_{TOT} = \sum_{i=1}^{J} d_i$

Step 4. Determine the relative weight of the i^{th} neighbor. : $w_i = \frac{1}{J-1}[1 - \frac{d_i}{d_{TOT}}]$

Step 5. Sum up each weight w_i that has the same output class in $w_{class_1}, \ldots, w_{class_c}$.

(n: The total number of output classes)

Step 6. Identify the class that has the highest sum of weights w_{class} and calculate the certainty-percentage p.: $p = (w_{class} - \frac{1}{n}) * \frac{n}{n-1} * 100$

Step 7. Separate the cases inside the grey-zone area as "undecided" and classify the rest of the cases.

Step 8. Interact with the user and reflect the final decision of the user in GCBR.

Figure 7. The algorithm of GCBR

4. Prototype of GCBR

In this chapter, the prototype of GCBR and how it works are described. The prototype of GCBR is implemented by JAVA. Figure 8 shows the initial interface of GCBR. The user sets the size of training, validation and test datasets then provides the information about the relationship between certainty-percentage and accuracy to the user. The screen providing this information is presented in Figure9. After that, the user can either directly decide a cut-off certainty-percentage or make the decision after simulating how much cut-off certainty-percentage minimizes the total cost in a given circumstance. If the "Simulation" button is pushed in Figure 8, the new interface pops up like Figure 10. This interface helps the user to simulate the desirable grey-zone in terms of cost concerns. GCBR gets the information about the user's recognized costs for each of the three cases – correctly classified, misclassified, and undecided cases and then suggests the desirable cut-off certainty-percentage for the user for minimizing the total cost in this state. Whether the user admits this suggestion or not is up to her/him. After the user sets the degree of the grey-zone and executes GCBR in Figure 10, the calculated results for the test dataset are returned as in Figure 11. However, the target cases less than the designated cut-off certainty-percentage are not classified into any group, thus the user has to make a decision for those cases. Finally, GCBR pops up a new interface as in Figure 12 for helping the user decide the result of the target cases classified as "undecided". The new interface gives the information on which previous similar cases are retrieved for each "undecided" target case and lets the user add the other cases or remove the neighbors. Even though GCBR helps the user by providing this function, the user makes the final decision. The upper side of Figure 12 is for supporting the user in making decisions and the bottom side is for reflecting the user's final decision to the result.

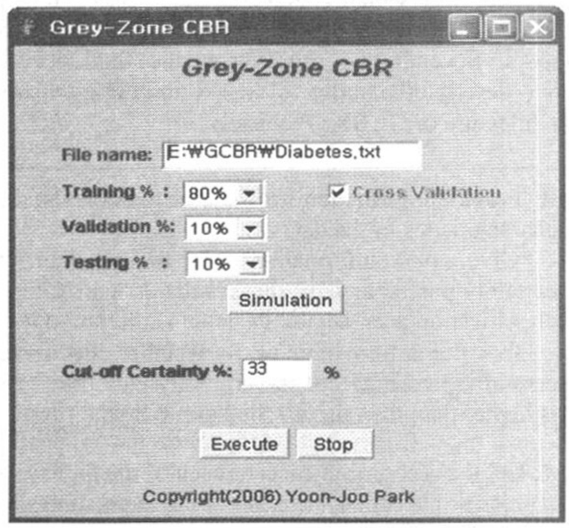

Figure 8. Initial interface of GCBR

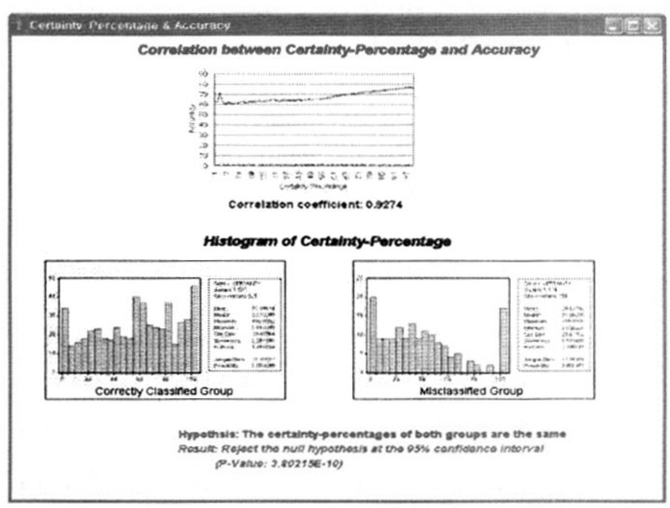

Figure 9. Providing information to for the user

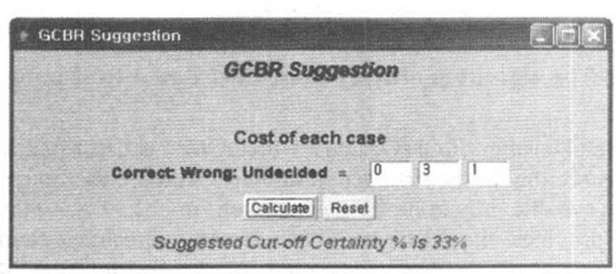

Figure 10. Interaction with the user concerning the cost of each case

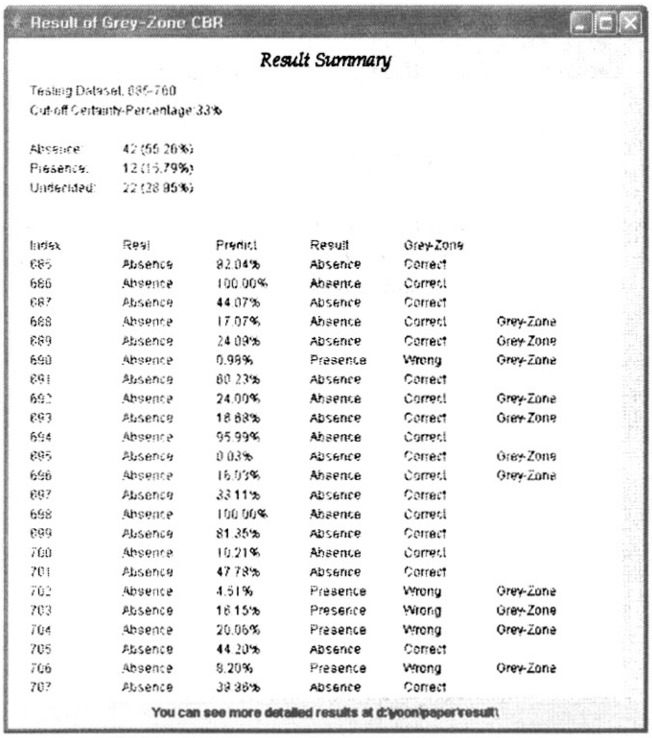

Figure 11. Result summary of GCBR

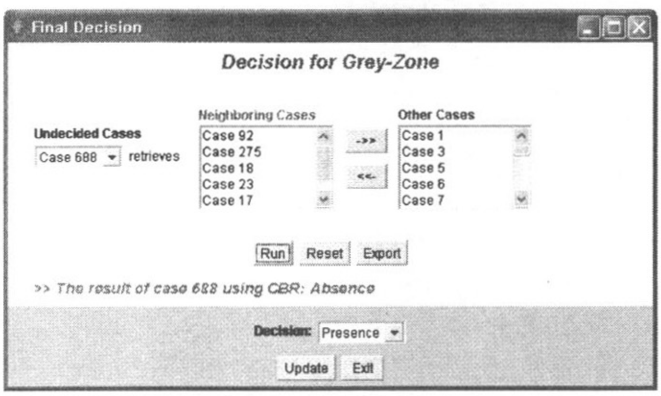

Figure 12. Interface for the undecided cases

5. SUMMARY AND CONCLUSIONS

This article suggests a new interactive CBR model called Grey-Zone Case Based Reasoning (GCBR) that can make a decision considering the cases near the cut-off point by interactive communication with users. GCBR classifies results automatically for the cases outside of the grey-zone; however it communicates with users to make a decision for the cases inside of the grey-zone by ensuring users understands that the characteristics the dataset and reflects the user's opinion.

In order to make an effective GCBR model we perform preliminary statistical analysis to figure out the correlation between certainty-percentage and accuracy. Also we perform the t-test to verify that the average certainty-percentage of the correctly classified group is significantly different from those of the misclassified group using diabetes dataset. Through these analyses, we identify that the validity of the assumptions for the dataset. GCBR is constructed based on these analyses. GCBR provides users with effective information for helping make decisions, understanding the user's needs and reflecting the user's opinion in the model. We introduce the architecture of GCBR that provides more deliberate and customized results to users and we also implement the prototype in this study.

We expect that this effort will help increase classification accuracy for the cases adjacent to the cut-off point and encourage the practical use of CBR in many real life areas.

References

Aamodt, A., E. Plaza (1994) Case-based reasoning: foundational issues, methodological variations, and system approaches, AI communications: the European journal on artificial intelligence, 7(1), 39-59.

Aha, D.W., A.B. Leonard, M.-A. Hector (2001) Conversational case-based reasoning, Applied Intelligence, 14(1), 9-32.

Aha, D.W., T. Maney, L.A. Breslow (1998) Supporting dialogue inferencing in conversational case-based reasoning, Lecture notes in computer science, 1488, 262-273.

Aktas, M.S., M. Pierce, G. C. Fox, D. Leake (2004) A Web based conversational case-based recommender system for ontology aided metadata discovery, Grid Computing, Proceedings. Fifth IEEE/ACM International Workshop on Publication Date: 8 Nov. 2004, ISSN: 1550-5510 ISBN: 0-7695-2256-4, 69- 75.

Blake, C.L., C.J. Merz (1998) UCI repository of machine learning database. Department of information and computer science, University of California, Irvine, CA (http://www.ics.uci.edu/~mlearn/MLRepository.html).

Chun, S.-H, Y.-J. Park (2005) Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction, Expert system with application, 28(3), 435-443.

Goker, M. H., C.A. Thompson (2000) Personalized conversational case-based recommendation, Lecture notes in computer science, 1898, 99-111.

Jurisica, I., Janice G., John M. (2000) Incremental iterative retrieval and browsing for efficient conversational CBR systems, Applied intelligence, 12(3), 251-268.

Leake, D.B., D.C. Wilson (2001) A case-based framework for interactive capture and reuse of design knowledge, Applied Intelligence, 14(1), 77-94.

Mchserry, D (2001) Interactive case-based reasoning in sequential diagnosis, Applied Intelligence, 14(1), 65-76.

Mekenna, E., B. Smyth (2001) An interactive visualisation tool for case-based reasoners, Applied Intelligence, 14(1), 95-114.

Park, Y.-J., B.-C. Kim, S.-H. Chun (2006) New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis, Expert systems, 23(1), 2-20

Simazu, H., Akihiro S., Katsumi N. (2001) ExpertGuide: A conversational case-based reasoning tool for developing mentors in knowledge spaces, Applied intelligence, 14(1), 33-48.

Yang, Q., J. Wu (2001) Enhancing the effectiveness of interactive case-based reasoning with clustering and decision forests, Applied Intelligence, 14(1), 49-64.