Implementation of Case-Based Reasoning System For Abalone

Lei Mon Ko
Computer University(Mandalay)
lmk287@gmail.com

Abstract

CBR (Case Based Reasoning) is an Artificial Intelligence methodology that provides the foundations for technology of intelligent systems. CBR receives increasing attention within the AI community.CBR is an analogical reasoning method providing both a methodology for problem solving and a cognitive model of people.Case-based reasoning is a recent approach to problem solving and learning that has got a lot of attention over the last few years. A new problem is solved by finding a similar past case, and reusing it in the new problem situation. The basic idea o f CBR is similar problem have similar solutions. The initial purpose of our system is to develop a case-based system where a new case could be quickly compared to the numerous cases in the databases. In our system. we provide the Abalone (sea ear) datasets to implement the system. If the user enters the attributes of abalone ,the system will display the rings of abalone. By adding 1.5 to the result ,the user will get the age of abalone. The system uses nearest-neighbor approach for case retrieval.ID3 algorithm is used in case adaptation.

1. Introduction

Abalone is a type of mollusc that grows predominantly in the waters in and around California .It is also known as sea-ear because of shape of their shell. Since discovering the beauty of abalone, people have used its shimmering shell for personal adornment items such as buttons and pendants. Abalone is known as "Mother of Pearl". Abalone shell consists of variety of elements including calcium ,iron ,magnesium ,sodium ,and silicon and several chemical compounds including aspartic acid and glutamic acid. Abalone shell can be used in medicine for liver diseases.

Case-Based Reasoning (CBR) is a popular reasoning methodology for decision support systems because its reasoning is based largely on case knowledge that may already be available in a database. When a new problem is presented to a CBR system, it first retrieves cases with similar problem descriptions from the case-base. The solutions in these retrieved cases are used to propose a solution for the new problem. It may be necessary to adapt the proposed solution to take account of differences between the new problem and the retrieved problems. In addition to returning the proposed solution as the answer to the new problem, it is common

to review the new problem and its solution, and perhaps to retain this problem-solution pairs a new case in the case-base. In this system, if the user input the attributes of Abalone , the system will output the age of Abalone. In section [2] we describe the related work of the system. In section [3] ,we describe the background theory of our system. Section [4] is the proposed system architecture of the system. Section 5 briefly describe the performance evaluation of the system.

2. Related Work

In [1] ,Case –based expert system is developed for supporting diagnosis of heart diseases: Mitral stenosis ,left-sided heart failure ,stable angina pectoris and essential hypertension. They implement two retrieval strategies :Induction and nearest neighbor approaches. The system has trained set of 42 cases for Egyptian cardiac patients. Each case is concerned with demographic and clinical data.

In [2],this paper research the retrieval algorithms for case in Case-based reasoning. They compare three retrieval algorithms: Voronoj-Inspired Quantitative Retrieval Algorithm, Tree-Hash Qualitative Retrieval Algorithm and Combined Quantitative and Qualitative Retrieval Algorithm.

In[5],the paper describes that every year Russia has more intoxication cases than any other country in Europe. Therefore it is reasonable to use valuable experience of the best Russian toxicologists. They describe an approach for developing knowledge-based medical decision support systems based on the rather new technology of case-based reasoning.

3. Background Theory

CBR presents a foundation for a new technology of building intelligent computer –aided diagnoses system. Case-Based Reasoning process as a cyclical process comprising of the four REs:[2]

1.Retrieve the similar cases

2.Reuse the case to attempt to solve the problem

3.Revise the proposed solution

4.Retain the modified solution as a new case

Figure 1 show the cycling of CBR. In CBR cycle, a new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case and retaining the new experience by incorporating it into the existing database.

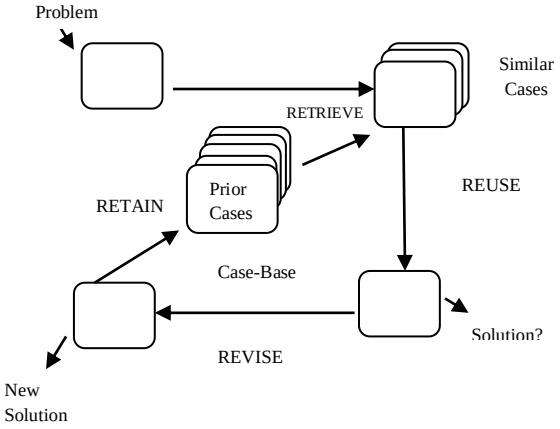

Figure 1: Case based Reasoning Cycle

3.1K-Nearest Neighbor Classifier

In retrieving the case, we use K-NN algorithm. In that algorithm ,closeness is defined in terms of Euclidean distance, where the Euclidean distance between two points ,$X(x_1,x_2,...,x_n)$ and $Y(y_1,y_2,...y_n)$ is

$$d(X,Y)= \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \qquad \text{Equ(1)}$$

3.2 Decision Tree Induction

In adapting the solution, the system use the ID3 algorithm .In this algorithm , information gain measure is used to select the test attribute at each node in the tree.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i=1,…..,m).

Let S_i be the number of samples of S in class C_i, The expected information needed to classify a given sample is given by

$$I(S_{1j},S_{2j},...,S_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2(p_{ij}) \text{ Equ(2)}$$

Where pi is the probability that an arbitrary sample belongs to class Ci and is estimated by si/s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct values,$\{a_1,a_2,..a_v\}$.Attribute A can be used to partition S into v subsets,$\{S_1,S_2,\ldots,S_v\}$,where S_j contains those samples in S that have value a of a. If A were selected as the test attribute(i.e., the best attribute for splitting),then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j}+...+S_{mj}}{S} I(S_{1j},S_{2j},...,S_{mj}) -$$

Equ(3)

The term $\frac{s_{1j}+...+s_{mj}}{s}$ acts as the weight of the jet subset and is the number of samples in the subset divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions .Note that for a given subset S_j,

$$I(S_{1j},S_{2j},...,S_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2(p_{ij})$$

Where $p_{ij} = \frac{s_{ij}}{|S_j|}$ and is the probability that a sample in S_j belongs to class C_i. The encoding information that would be gained by branching on A is

Gain (A) = I ($s_1,s_2,.....,s_m$) – E(A)-equ(4)

In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

3.2.1Adaptation the solution

If there is identical case in the database , the system directly display the result. Otherwise, the system needs to adapt the solution. The system use ID3 algorithm to adapt the solution. The decision tree algorithm used in this paper is summarized as follows:

Algorithm : Generate_decision_tree

Input : The trainin samples,samples,represented by discrete-valued attributes;theset of candidates attributes,attribute-list.

Output : A decision tree

Method :

(1)create a node N;

(2)if samples are all of the same class, C then

(3) return N as a leaf node labeled with the most common class in samples;

(4)if attribute-list is empty then

(5) Return N as a leaf node labeled with the most common class in samples;

(6) select test-attribute, the attribute among attribute-list with the highest information gain;

(7) label node N with test-attribute;

(8) for each known value a_i of test-attribute

(9) grow a branch from node N for the condition test-attribute=a_i;

(10) let s_i be the set of samples in samples for which test-attribute=a_i;

(11) if s_i is empty then

(12) attach a leaf labeled with the most common class in samples;

(13) else attach the node returned by Generate_decision_tree;

4.Proposed System Architecture

The figure2 shows the architecture of the proposed system. When the user inputs new data to the system, the user needs to enter the value of K. We show the example with 10 datasets. In one case, normally there are 8 attributes: Rings, Length (measurement of shell), Diameter (perpendicular to length), Height (with meat in shell), Whole Weight (Whole abalone), Shucked Weight (Weight of meat), Viscera Weight (gut weight), and Shell Weight (After being dried). But in this example, we show with 6 attributes as shown in Table (4.1).

After that the system will process the KNN algorithm. In this phase, the system can retrieve either the exact match with the new case or the nearest match which has the greatest similarity with the new case. If the exact match is found, the system gives the exact case solution to the user directly. Otherwise, the system will retrieve the general solutions according to value K. If the user enters the value of k=3,the system shows the 3 nearest neighbours of the input case as shown in Table (4.2) by using the Equation (1). In this table, F is female and M is Male. The class label is age.

Then it revises that solution by using adaptation rules in order to display the appropriate solution to the user. To calculate the adaptation rules, firstly it needs to calculate the expected information of the class label by using Equation (2).Let s1 be 10 and s2 be 7.I(s1,s2) is 0.917.And then we need to calculate the entropy of each attribute by using Equation (3) and calculate Gain by using Equation (4).

For ”sex” E(sex)=0.917 and G(sex)=0

For ”length” E(length)=0.667 G(length)=0.25

For “Diameter” E(diameter)=0 G(Diameter)=0.917

For “height” E(height)=0.645 G(height)=0.272

For ”whole weight” E(WW)=0.667 G(WW)=0.25

For”Shucked Weight” E(SW)=0.667 G(SW)=0.25

So the diameter is assumed as root node of the decision tree.

Roughly the system generates rules as follows:

If diameter is” 0.34” then Rings=”10”

Else If diameter is”0.31`” then Rings=”7”

Else If diameter is”0.36” then Rings=”10”.

Table 4.1 Dataset with 6 attributes

Sex	Length	Diameter	Height	Whole weight	Shucked Weight	Rings
M	0.44	0.365	0.125	0.516	0.2155	10
F	0.44	0.34	0.1	0.451	0.188	10
M	0.405	0.31	0.1	0.45	0.173	7
F	0.47	0.355	0.1	0.4755	0.1675	10
M	0.4	0.32	0.095	0.303	0.1335	7
M	0.425	0.325	0.095	0.3785	0.1705	7
M	0.425	0.35	0.105	0.393	0.13	9
F	0.325	0.26	0.09	0.1915	0.085	7
F	0.405	0.325	0.11	0.3555	0.1511	9
F	0.47	0.375	0.125	0.5615	0.252	10

Table 4.2 3-Nearest Neighbours of new instance

Sex	Length	Diameter	Height	Whole Weight	Shucked Weight	Rings
F	0.44	0.34	0.1	0.45	0.19	10
F	0.44	0.31	0.1	0.45	0.17	7
F	0.47	0.36	0.1	0.39	0.17	10

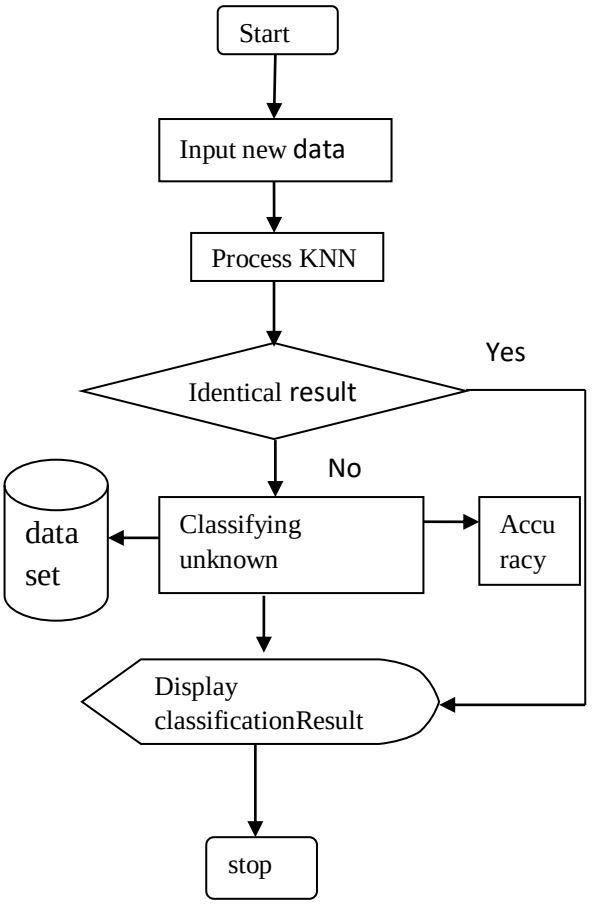

Figure2.The Proposed System Architecture

5. Evaluation of Performance

In calculating the accuracy, we use the hold- out method .In the holdout method, the given data are randomly partitioned into two independents sets; training set and testing sets. Typically , two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimate with the test set. The estimate is pessimistic since only a portion of the initial data is used to derive the classifier. Random subsampling is a variation of the holdout method in which the holdout method is repeated k times. The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration. In this system , the accuracy is evaluated depends on decision tree induction.

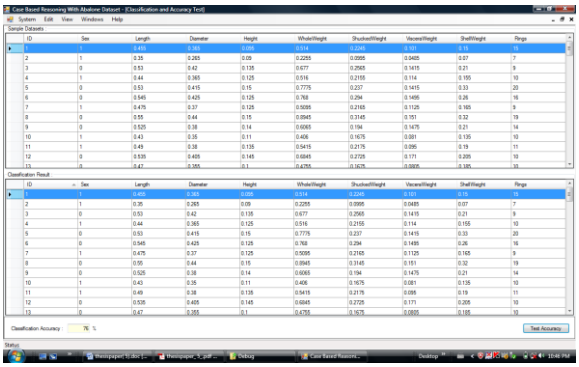

Figure3.Accuracy of the system

6. Conclusion

Case-Based Reasoning puts forward a paradigmatic way to attack AI issues, namely problem solving, learning, usage of general and specific knowledge ,combining different reasoning methods. Case-based reasoning emphasizes problem solving and learning from experiences.CBR approach appears to have some advantages concerning system development if compared with other knowledge-based method. The proposed system will identify the age of Abalone by entering the attributes of it. By knowing the age of abalone, the user can identify in what age of the abalone, how the organs of it develop. The researcher can decide it was ready for production of personal adornment items. And at the farm of abalone, the user can see how abalone is growing within a duration. The proposed system can only be used for Abalone datasets. These datasets are collected from UCI machine learning repository. The system can be extended by using other classification method such as Fuzzy set , Naive Bayesian etc.

7. References

[1]Abdel-Badeeh M.Salem"A Case –based Expert System for supporting Diagnosis Of Heart Diseases".

[2]Alamodt & Plaza."Case-based Reasoning Foundational Issues,Methodological Variations and System Approaches,AI Communication Vol.7 Nr.1 March 1994

[3]I .Watson,F.Marir,"Case-based Reasoning:A Review", Vol 9,No 4,1994

[4] Jiawei Han,Micheline kamber"Data Mining:Concepts and Techniques"

[5Klaus-Dieter Althoff,Ralph Begmann"Case-based Reasoning For Medical Decision Support Tasks:The INRECA Approach

[6Kolodner,J.L., ‘Case-Based Reasoning,”Morgan Kaufmann,1993.

[7]Padraig Cunningham and Sarah Jane Delany“K Nearest Neighbour Classifiers”Technical Report UCD-CSI-2007-4March 27,2007

[8]www.ucimachinelearningrepository.com