Implying p-Cure algorithm in case retrieval stage of the case-based reasoning

Apurva Mittal[1], Dr. Surjeet Dalal[2]

[1]Student, M. Tech, ESEAR, Ambala
[2]Associate Professor, Dept. of CSE, E-Max group of Institutions, Ambala

***Abstract*— The previous methods tried to find a compact representation of the data that can be used for future prediction. In case-based reasoning, the training examples - the cases - are stored and accessed to solve a new problem. To get a prediction for a new example, those cases that are similar, or close to, the new example are used to predict the value of the target features of the new example. This is at one extreme of the learning problem where, unlike decision trees and neural networks, relatively little work must be done offline, and virtually all of the work is performed at query time. Cure can detect cluster with non-spherical shape and wide variance in size using a set of representative points for each cluster. Cure has a good execution time in presence of large database using random sampling and partitioning methods and also works well when the database contains outliers. In this paper we implies p-Cure algorithm to enhance the performance of the case-based reasoning system**.

***Keywords*— Case-based reasoning, Case retrieval, Clustering, Cure algorithm**.

I. Introduction

The case-based reasoning is a problem-solving loom that replicates the human being problem-solving conduct. In this approach, the trouble is being solved out on origin of precedent experiences gained from throughout the process of solving the problem in the earlier period. In case of multifaceted system, it is extremely complicated to devise the circumstances with domain rules. Other shortcoming is that the rules necessitate supplementary input information than is characteristically accessible, because of imperfect problem specifications or because the knowledge needed is simply not available at problem-solving time. But in case of CBR approach, if general knowledge is not sufficient because of too many exceptions, or when new solutions can be derived from old solutions supplementary without difficulty than from scuff, then on foundation of precedent experiences, the problem is being solved.

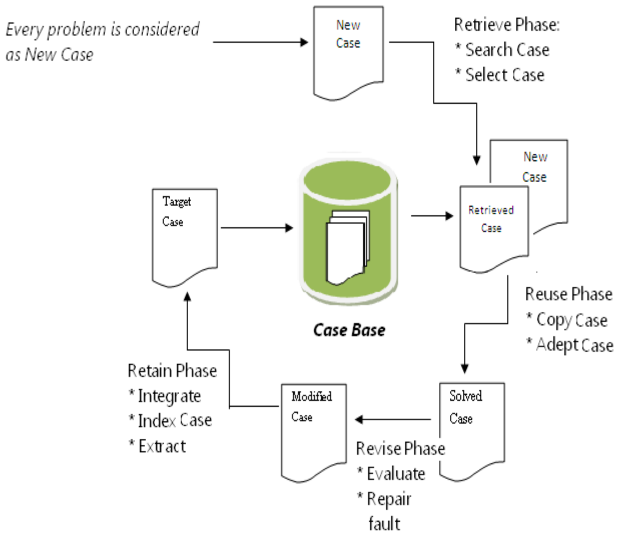

Figure 1: Case Based reasoning Cycle.

The case based reasoning involves four phases in the dilemma solving as given below:

- Retrieval phase.
- Reuse phase
- Revise phase
- Retain phase

Every problem pattern & its pledge are stored in form of the cases. It maintains the accumulation of the cases that is recognized as the case base. In this system, every problem is measured as the new-fangled case. In the retrieve phase according the new case, estimated solution case is being searched from the case base & preferred. After the assortment of the case, that case is adapted with the innovative case. It generates the resolved case. Now the solved case is evaluated in the revise phase & the faults in that case are being repaired. Now customized case is the resolution of the problem. This solution is stored in the case with appropriate index. This action is compulsory for extracting the cases very competently & rapid access to the cases in prospect.

Case-Based Reasoning is not constrained to the salvage of the experience. Another very important feature of case based reasoning is its coupling to learning. As the human beings learns from the precedent familiarity, the case base reasoning supports erudition from the history experience. Learning in CBR occurs as a usual by contraption of problem solving. When a problem is lucratively solved, the experience is retained in order to resolve comparable problems in the potential. When an attempt to solve a problem fails, the motive for failure is identified and remembered in order to shun the identical blunder in the future.

I.CASE RETRIEVAL PHASE

Retrieval is a major research area in CBR. Case retrieval is the process of extracting those cases within a case base that are the closest to the current case. It is always required to extract the best cases from the casebase and for that a better and effective selection criteria is needed. The selection criteria determine which is the best case to retrieve, by determining how close the current case is to the cases stored.

Many traditional retrieval techniques so far being used in the CBR system are k-nearest neighbors (k-NN), decision trees, inductive approaches, Knowledge-guided approaches. These techniques involve developing a similarity metric that allows closeness (i.e., similarity) among cases to be measured. The choice of retrieval techniques in CBR applications requires experience and experimentation

The most commonly used technique in the CBR system is k-nearest neighbors (k-NN).this technique is based on the concept of weighted features. Features that are considered more important in a problem-solving situation may have their importance denoted by weighting them more heavily in the case-matching process. A case that matches the present case on n features will be retrieved rather than a case that matches on only k features, where $k < n$.

Although this technique is simple but its performance declines when the size of case base increases.The main issue and problem faced in the retrieval phase of CBR system is that system becomes slow when size of case bases grows. To resolve this problem we are going to use some clustering technique with thy K-NN algorithm which can overcome the limitation of the K-NN algorithm without compromising its retrieval speed.

II.RELATED WORK

In the past, work has been done in many fields of data mining, document retrieval, information retrieval and in many applications of image processing. Different Clustering based techniques are being applied to the applications to improve the performance for their successful execution.

(Sadeghi and Teshnehlab 2008) presented a new Ant Clustering algorithm based on case based reasoning Every ant has a case base which is updated iteratively by the process of CBR. Each ant can use its case base to find best places for dropping its load. They gave a mechanism for taking the advantage of knowledge of other ants' case bases by the process of cooperation. They used a probabilistic way for data items that are scattered randomly over the 2-D space to be picked up, transported and dropped by the agents for which they used probability and density functions for picking and dropping operations. Those ants that were unsuccessful in dropping their loads can give their loads to other unladen ants in their neighborhood, and those that were unsuccessful in picking up an object can take the loads of other laden ants in their neighborhood. If two neighboring ants both were laden they can exchange their loads. This process speeded up the process of clustering and prevented from producing too many small clusters which is one of main drawbacks of ant based clustering. The results demonstrated better performance in terms of accuracy and compactness of generated clusters than previous approaches of ant based clustering.

(Smiti and Elouedi 2013) propose a novel case base competence model based on Mahalanobis distance and a clustering technique named DBSCAN-GM. The advantage of this newly proposed model is its high accuracy for predicting competence. In addition, it is not sensitive to noisy cases and it takes account the situation of the distributed case-base. They remarked the model could be very useful in future CBR research in fields such as the development of new policies for maintaining the case base.

III. CURE ALGORITHM

To deal with the large casebase CBR systems often uses clustering techniques to classify the similar cases in different groups. Cure clustering is considered very efficient clustering algorithm as it is robust to outliers and identifies clusters with non-spherical shapes, and wide variances in size. Cure can handle high dimensional data without sacrificing its clustering quality.

Combining the cure clustering technique with KNN algorithm improves the performance performance of the CBR system to a great extent. The idea of the modified algorithm works as:-

Cure algorithm random sample the data set and perform partitioning. For each cluster, choose a constant no. c of well scattered points. Point denotes the cases. Shrink scattered points towards the centroid of the cluster by a shrink factor α where $0\leq \alpha \leq 1$. These scattered points are used as representatives of the cluster. Distance between the two clusters is computed by calculating the distance between the closest pair of representative points from each cluster. Merge the closest pair of clusters at each iteration until M clusters are found. Update the distance matrix if M clusters are not found otherwise calculate Euclidian distance between the query-instance and all set of points in M clusters. Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance. Then the similarity is calculated between stored cases and new input case based on weighted feature.

IV. PCURE EXECUTION FOR FAST RETRIEVAL IN DCBR

Although random sampling and partitioning allows CURE to handle large datasets the algorithm is not applicable to today's huge databases because of its quadratic time complexity. The above technique proves efficient when working with single processor CBR system.

In Distributed case-based reasoning system (DCBR), where problems are solved by the combined effort of multiple, independent CBR agents, efficiency can be improved by distributing problem solving effort across multiple independent agents operating in parallel. The problem is distributed among shared memory multiprocessors. Retrieval from very huge databases of the Distributed case-based reasoning systems is a very challenging task. A parallel implementation of Cure here can prove more efficient hierarchical data clustering algorithm.

In Distributed CBR systems problem solving experience and responsibility is distributed among multiple CBR systems or agents. Individual agents solve their own problems that fall in their category but draw on the experience of other agents for problems that do not. In DCBR systems multiple case bases are maintained. There are two types of DCBR systems:-

Single-agent multiple case-bases case-based reasoning systems – where there is single problem solving agent access these multiple case-bases.

Multiple-agent multiple case-bases case-based reasoning systems—where problem solving effort is distributed across multiple CBR agents. Each agent solves each target problem locally. When a given agent, does not have the experience to solve the target problem then it cooperates with other agents by borrowing their problem solving experience.

The agents have the same CBR capabilities, but differ in their problem solving experiences. An agent case-base, may focus on specific regions of the target problem space, and thus each agent may differ in their problem solving coverage, thereby allowing different agents to solve different problem types.

For fast retrieval from huge databases, a parallel implementation of Cure uses a linear array of records that keeps information about the size, centroid and representative points of each cluster. Parallelization is possible because there are no data dependencies. Here, we are improving the merging procedure of the clusters. On merging two clusters we store the information of the resulted cluster in the entry of the first cluster and simply invalidate the second one. In this each cluster have to maintain per cluster information about the index of the closest cluster and the minimum distance to it. The algorithm searches closest neighbor with the smallest index in the entry of the given cluster. This process consumes less memory and speeds up the sequential algorithm.

V. RESULTS AND CONCLUSION

Three sections are discussed in this paper- a simple KNN based CBR, a modified approach of KNN using cure clustering in CBR and its parallel implementation in DCBR. Our motive is to fast the retrieval process of the CBR systems when working on large data sets.

In the complexity of classical approach, the number of points drastically increase the complexity as the distance will also affect the overhead on the algorithmic approach. The Time Complexity of CLASSICAL K NEAREST NEIGHBOR is O (d n2). Where d is the Distance and n is the Number of Elements.

For increasing or large data sets, the performance of KNN degrades abruptly as the major factors are number of points (n) and the distance. This issue is required to be removed by integration of any other inner algorithmic approach. So, The Cure clustering is applied to the large data sets. Cure can detect cluster with non-spherical shape and wide variance in size using a set of representative points for each cluster. Cure has a good execution time in presence of large database using random sampling and partitioning methods and also works well when the database contains outliers.

The time complexity calculated after combining it with KNN approach is O (n log n). The biggest advantage of using cure with KNN algorithm is that it provides good quality of clustering

REFERENCES

[1] J. Kolodner, "Case Based Reasoning. Morgan Kaufmann," San Francisco, 1993.

[2] S. W. Changchien, M. C. Lin, "Design and implementation of case based reasoning system for marketing plans," Expert Systems with Applications, vol. 28, pp. 43–53, 2005.

[3] J. Sebestyénová, "Case-based Reasoning in Agent-based Decision Support System", *Acta Polytechnica Hungarica*, Vol. 4, No. 1, 2007, pp.127-138.

[4] Chen Juan, Yang Ying. Based on case-based reasoning teaching cases knowledge management system design [J]. Computer and Information Technology, 2010, 18(3): 57-59.

[5] Kevin Vogts, Nigel Pope. Generating Compact Rough Cluster Descriptions Using an Evolutionary Algorithm [J], Lecture Notes in Computer Science, 2004, 3103: 1332-1333.

[6] Hichem Frigui. SyMP: An Efficient Clustering Approach to Identify Clusters of Arbitrary Shapes in Large Data Sets [C], In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (SICKDD), 2002, pages 507-512.

[7] Yong-Bin Kang, Shonali Krishnaswamy, and Arkady Zaslavsky, "A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge", IEEE transactions on cybernetics, vol. 44, no. 4, april 2014 pp. 473-487.

[8] Anantaporn Srisawat, Tanasanee Phienthrakul, Boonserm Kijsirikul. SV-kNNC: An Algorithm for Improving the Efficiency of k-Nearest Neighbor [J], Lecture Notes in Computer Science, 2006, 4099: pp.975-978.

[9] Xipeng Qiu, Lide Wu. Nearest Neighbor Discriminant Analysis [J], International Journal of Pattern Recognition and Artificial Intelligence, 2006, 20(8): 1245-1259.

[10] G. R. Beddoe and S. Petrovic, "Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering," Eur. J. Oper. Res., vol. 175, no. 2, pp. 649–671, 2006.

[11] K. Bradley and B. Smyth, "Personalized information ordering: A case study in online recruitment," Knowl.-Based Syst., vol. 16, nos. 5–6, pp. 269–275, 2003.

[12] C. M. Vong, P. K. Wong, and W. F. Ip, "Case-based classification system with clustering for automotive engine spark ignition diagnosis," in Proc. 9th Int. Conf. Comput. Inf. Sci., Aug. 2010, pp. 17–22.

[13] F. Azuaje, W. Dubitzky, N. Black, and K. Adamson, "Discovering relevance knowledge in data: A growing cell structures approach," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 30, no. 3, pp. 448–460, Jun. 2000.