Performance evaluation of fuzzy clustered case based reasoning

Houssam Zitan , Zakaria Amgrout
Supervised by Pr. Khalid Jebbari
Faculty of science and technology tangier FSTT

March 17, 2025

Abstract

Case-Based Reasoning (CBR) is a machine learning technique that solves new problems using past cases stored in a case base. While this improves problem-solving capability, continuous case-base growth creates performance challenges. The key issues are managing this growth and maintaining case relevance without bottlenecks. Various solutions exist, and this work proposes a fuzzy clustering-based knowledge maintenance approach that enhances performance without compromising effectiveness. The proposed method is evaluated on different case bases and compared with conventional CBR.

Introduction

Case-Based Reasoning (CBR) is a lazy learning approach that solves new problems using past cases stored in a case base. While its incremental growth enhances problem-solving capability, it also introduces performance bottlenecks, particularly in case matching. Clustered CBR has been proposed to address efficiency issues, but conventional clustering creates rigid boundaries, limiting knowledge retrieval. Cases near cluster boundaries may have relevance to neighboring clusters, which rigid clustering fails to capture. To overcome this, a fuzzy clustering-based approach is proposed, allowing cases to belong to multiple clusters based on membership degrees. This improves knowledge acquisition and problem-solving efficiency by reducing information loss. The proposed hybrid model integrates fuzzy logic, clustering, and CBR to enhance accuracy, recall, precision, and time complexity. Experimental evaluation demonstrates its superiority over conventional CBR in multiple problem domains.

1 Case-based reasoning

Case-Based Reasoning (CBR) is inspired by human problem-solving behavior, using past solved cases to address new problems. These cases are stored in a knowledge base, where similar cases are retrieved, adapted, and retained for future use, enriching knowledge over time. CBR is widely applied across various domains, including e-Commerce, medical diagnosis, fault detection, law, and software quality prediction. It offers key advantages such as minimal preprocessing, flexible adaptation, and low learning requirements. The CBR process consists of four stages, known as 4-Rs, which guide problem solving efficiently: Case retrieval, Case reuse, Case revision and Case retention.

Limitations of conventional Case based reasoning System

Conventional CBR systems compare new problems with all existing cases, regardless of relevance, leading to inefficiencies such as:

- Unnecessary comparisons with irrelevant cases.
- Increasing computational cost with each new problem.

Over time, the growing case base results in longer problem-solving times, which causes efficiency bottlenecks. When this occurs, traditional case-base organization strategies become ineffective, necessitating advanced knowledge maintenance approaches.

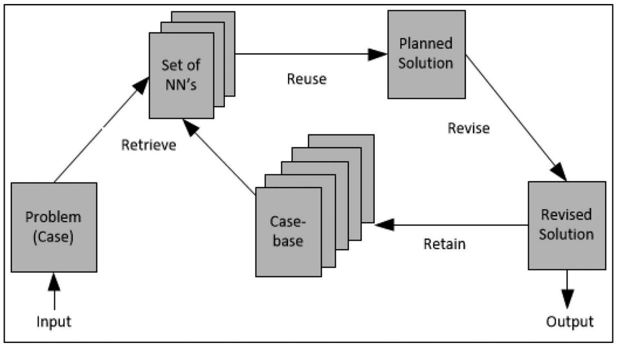


Figure 1: Conventional case-based reasoning.

Clustering and Case-Based Reasoning

As data size grows, extracting meaningful patterns becomes challenging. Clustering, an unsupervised learning approach, groups similar data into clusters, but determining the optimal number of clusters remains a research problem. Conventional clustering struggles with imbalanced data, leading to the development of clustering ensembles for improved performance. To address CBR inefficiencies, clustered CBR was introduced, where cases are grouped using k-means clustering. When a new problem arises, it is classified into the most relevant cluster, reducing unnecessary comparisons and computational costs. This approach enhances efficiency by confining case retrieval to a specific cluster, optimizing the knowledge-base maintenance process.

Limitations of clustered CBR

Clustered CBR improves efficiency by restricting case retrieval to similar cases, reducing computational costs. However, it has key limitations:

- As clusters grow large, time complexity may again become a concern.
- Rigid clustering assigns cases exclusively to one cluster, causing loss of crucial boundary information, potentially leading to less effective solutions.

These limitations create knowledge acquisition bottlenecks that ultimately affect the overall performance of the CBR process.

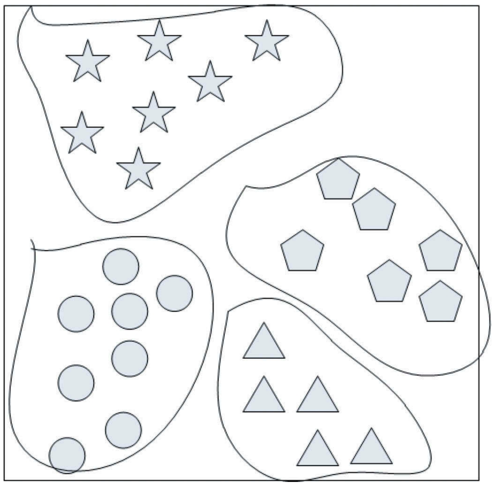


Figure 2: Clustered case-base with rigid boundaries.

2 Fuzzy clustered CBR

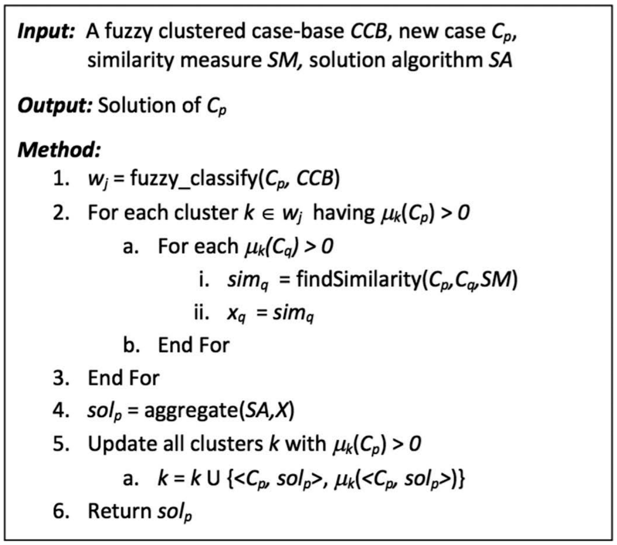
To overcome knowledge acquisition challenges in CBR, a fuzzy clustering approach is proposed. Unlike rigid clustering, this method maintains the knowledge base as overlapping fuzzy clusters, allowing cases to belong to multiple clusters based on their relevance. This reduces information loss and enhances decision-making. The approach includes two key algorithms:

- partitioning the case base into fuzzy clusters (**Algorithm 1**).
- solving new problems using fuzzy classification and CBR (**Algorithm 2**).

This methodology improves efficiency and adaptability in CBR-based problem-solving models.

Offline maintenance activity of fuzzy clusters

Algorithm 1 utilizes the fuzzy c-means clustering algorithm in a brute-force approach to partition the case base into an optimal number of fuzzy clusters. It begins by initializing accuracy at zero and clusters the case base into two groups in the first iteration. Using a leave-one-out (LOO) validation approach, classification and the CBR cycle are applied to each case, updating accuracy at each step. The process continues iteratively, determining the optimal number of clusters based on the highest CBR accuracy. Clustering stops when further division becomes ineffective, ensuring meaningful case distribution and efficient problem-solving.

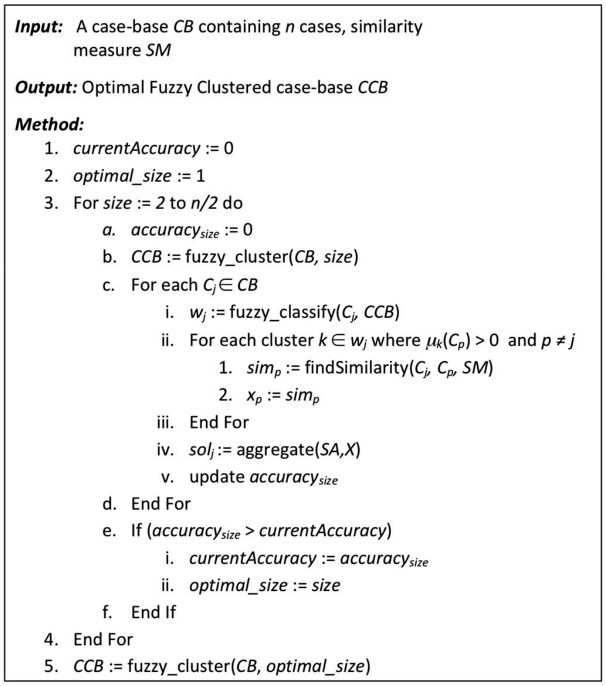


```
Input:  A fuzzy clustered case-base CCB, new case $C_p$,
        similarity measure SM, solution algorithm SA
Output: Solution of $C_p$
Method:
  1. $w_j$ = fuzzy_classify($C_p$, CCB)
  2. For each cluster $k \in w_j$ having $\mu_k(C_p) > 0$
       a. For each $\mu_k(C_q) > 0$
            i.  $sim_q$ = findSimilarity($C_p$,$C_q$,SM)
            ii. $x_q = sim_q$
       b. End For
  3. End For
  4. $sol_p$ = aggregate(SA,X)
  5. Update all clusters k with $\mu_k(C_p) > 0$
       a. $k = k \cup \{<C_p, sol_p>, \mu_k(<C_p, sol_p>)\}$
  6. Return $sol_p$
```

Figure 3: Algorithm-1 : partitioning the case-base into fuzzy clusters.

Online decision making activity using fuzzy clustered CBR

Algorithm 2 uses the optimally clustered case base to solve new problems. When a new case arrives, fuzzy classification identifies relevant clusters, confining the solution space to those with non-zero membership. Each relevant case is compared to the new case using a similarity measure, where similarity magnitude determines its weight in the solution computation. A weighted aggregation method then derives the final solution, which is stored in the case base along with its fuzzy membership vector, ensuring efficient knowledge retention and retrieval.



```
Input:  A case-base CB containing n cases, similarity
        measure SM
Output: Optimal Fuzzy Clustered case-base CCB
Method:
  1. currentAccuracy := 0
  2. optimal_size := 1
  3. For size := 2 to n/2 do
       a. $accuracy_{size}$ := 0
       b. CCB := fuzzy_cluster(CB, size)
       c. For each $C_j \in CB$
            i.   $w_j$ := fuzzy_classify($C_j$, CCB)
            ii.  For each cluster $k \in w_j$ where $\mu_k(C_p) > 0$ and $p \neq j$
                   1. $sim_p$ := findSimilarity($C_j$, $C_p$, SM)
                   2. $x_p := sim_p$
            iii. End For
            iv.  $sol_j$ := aggregate(SA,X)
            v.   update $accuracy_{size}$
       d. End For
       e. If ($accuracy_{size}$ > currentAccuracy)
            i.  currentAccuracy := $accuracy_{size}$
            ii. optimal_size := size
       f. End If
  4. End For
  5. CCB := fuzzy_cluster(CB, optimal_size)
```

Figure 4: Algorithm-2 : decision making process using fuzzy clustered CBR.

Nearest Neighbors and Representation

When a new problem is presented, the CBR system retrieves its nearest neighbors using the k-nearest neighbors algorithm. Each neighbor has a weight reflecting its contribution to the solution, influenced by its membership strength in different fuzzy clusters. The classification process assigns a membership vector to the new case, representing its relevance to multiple clusters. The final weight of each neighbor is determined using a fuzzy aggregation technique :

$$w_k = \max\left(\min(\mu_i^1, \mu_j^1), \ldots, \min(\mu_i^m, \mu_j^m)\right)$$

After computing the solution, it is fine-tuned and, if successful, stored in the case base along with its membership vector for future problem-solving.

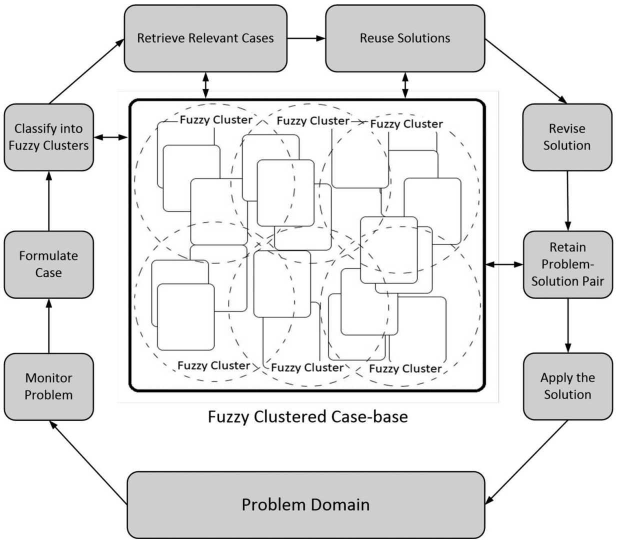


Figure 5: Fuzzy clustered case-base reasoning cycle.

3 Cases Studies

3.1 Autonomic Forest Fire Application (AFFA)

The FireForest dataset contains various attributes that help in analyzing forest fire occurrences. It includes geographical coordinates (X, Y) and temporal details (month, day) to pinpoint when and where fires may occur. Key meteorological factors are also provided, such as FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), and ISI (Initial Spread Index), which are indicators of fuel moisture and fire potential. Additionally, the dataset records weather conditions including temperature (temp), relative humidity (RH), wind speed, and rainfall (rain). The target variable, fire, is binary (0 or 1), indicating the absence or occurrence of a fire, respectively.

Algorithm 1 has been executed on a case-base of 517 cases and Algorithm 2 has been used to evaluate the performance of the proposed approach using LOO validation and accuracy.

The algorithms gave the best number of clusters as 9 using the Euclidean similarity metric, with LOO accuracy of 58.60%.

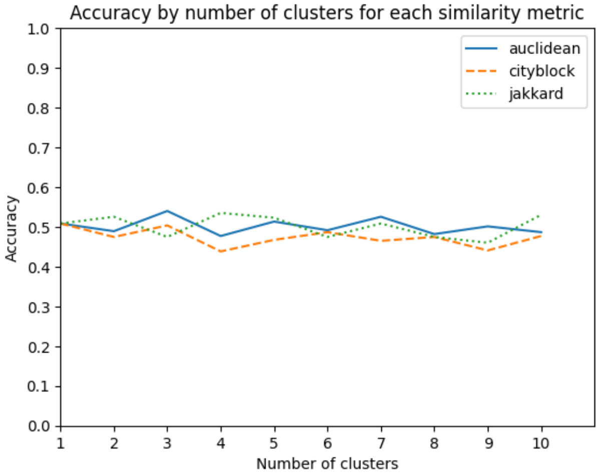


Figure 6: Accuracy behaviour on AFFA.

3.2 Case-base of Wisconsin breast cancer data

The Wisconsin Breast Cancer dataset is a popular dataset used for classifying breast cancer as malignant or benign based on various features of cell nuclei present in breast cancer biopsies. The dataset contains 569 instances and 30 features, including measurements of the cell's characteristics, such as radius, texture, smoothness, and compactness. These features are numerical and describe properties like the size and shape of the cell nuclei. The target variable in the dataset is binary, where 0 represents benign tumors and 1 represents malignant tumors.

The algorithm gave the best number of clusters as 21 using the Cityblock similarity metric, with LOO accuracy of 97.14%.

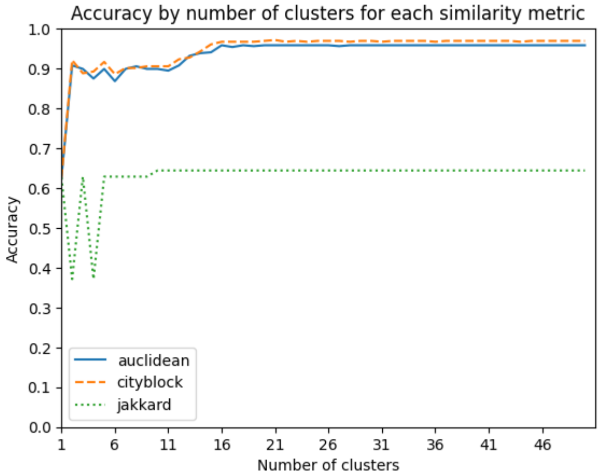


Figure 7: Accuracy behaviour on Breast cancer.

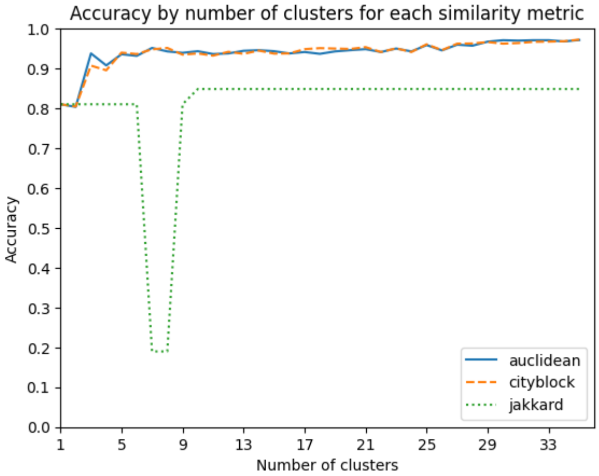


Figure 8: Accuracy behaviour on Skin Segmentation.

3.3 Skin detection

The Skin Segmentation dataset is a collection of data used for classifying pixels in images as either skin or non-skin.
The goal of the dataset is to help build machine learning models that can identify human skin in images or video frames, which has applications in fields like facial recognition, biometric security, and image processing.
This dataset contains features extracted from the color space of the image, such as RGB values, along with other color-based representations.
It has both labeled skin and non-skin regions, allowing for supervised learning. The dataset includes over 245,057 instances, where each instance corresponds to a pixel in an image, labeled as either 1 (skin) or 0 (non-skin).
The features represent the color characteristics that can help distinguish skin tones from other objects in the image.
The algorithm gave the best number of clusters as 15 using the Cityblock similarity metric, with LOO accuracy of 99.690%.

Testing results

Table 1 presents the performance of clustering on three different datasets, showing the best number of clusters and the corresponding accuracy achieved. For the Breast Cancer dataset, the optimal number of clusters was 21, resulting in an accuracy of 97.14%. The Forest Fire dataset had 9 clusters with an accuracy of 58.60%, while the Skin Segmentation dataset achieved the highest accuracy of 99.69% with 15 clusters. These results highlight the varying effectiveness of clustering across different datasets.

Case Base	Best Number Of Clusters	Accuracy
Breast Cancer	21	97.14%
Forest Fire	9	58.60%
Skin Segmentation	15	99.69%

Table 1: Performance of clustering on different datasets

Table 2 shows the best similarity measure used for clustering on three different datasets. For the Breast Cancer dataset, the ”cityblock” similarity measure was the most effective, while the ”euclidean” measure performed best for the Forest Fire dataset. The Skin Segmentation dataset also performed best with the "cityblock" similarity measure. These results indicate the varying impact of similarity measures on clustering performance across different datasets.

Case Base	Best Similarity Measure
Breast Cancer	cityblock
Forest Fire	euclidean
Skin Segmentation	cityblock

Table 2: Best Similarity Measure For Each Dataset

Table 3 compares the prediction running times with and without clustering for three datasets. For the Breast Cancer dataset, clustering reduced the running time from 0.4168s to 0.2243s. In the Forest Fire dataset, the time difference was smaller, with clustering taking 0.5215s compared to 0.5750s without it. The Skin Segmentation dataset showed the most significant improvement, where clustering reduced the time from 5.7018s to 1.0573s. These results indicate that clustering can substantially improve prediction efficiency, especially for larger datasets.

Case Base	With Clusters	No Cluster
Breast Cancer	0.2243s	0.4168s
Forest Fire	0.5215s	0.5750s
Skin Segmentation	1.0573s	5.7018s

Table 3: Prediction Running Time Comparison

Conclusion

To address knowledge acquisition and performance bottlenecks in conventional CBR, a hybrid approach partitions the case-base into an optimal number of fuzzy clusters using a brute-force search algorithm. Each case is assigned a fuzzy membership vector, allowing classification into multiple relevant clusters and reducing the search space for new problems. This fuzzy clustering minimizes information loss by enabling overlapping cluster memberships. The model was evaluated using three case studies, analyzing accuracy across different distance metrics. Results indicate that optimal clustering varies by dataset, and distance measure selection impacts decision-making. Future work aims to incorporate structural similarity for improved case retrieval.

References

[1] Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications, 7*(1), 39–59. https://doi.org/10.3233/AIC-1994-7104.

[2] Abbasi, S., Nejatian, S., Parvin, H., Rezaie, V., & Bagherifard, K. (2019, August). Clustering ensemble selection considering quality and diversity. *Artificial Intelligence Review, 52*(2), 1311–1340. https://doi.org/10.1007/s10462-018-9642-2.

[3] Abdel-Aziz, A., & Hüllermeier, E. (2015). Case base maintenance in preference-based CBR. *Lecture Notes in Computer Science, 9343*, 1–14. https://doi.org/10.1007/978-3-319-24586-7_1.

[4] Amini, S., Homayouni, S., Safari, A., & Darvishsefat, A. A. (2018). Object-based classification of hyperspectral data using random forest algorithm. *Geo-spatial Information Science, 21*(2), 127–138. https://doi.org/10.1080/10095020.2017.1399674.

[5] Bagherinia, A., Minaei-Bidgoli, B., Hossinzadeh, M., & Parvin, H. (2019, May). Elite fuzzy clustering ensemble based on clustering diversity and quality measures. *Applied Intelligence, 49*(5), 1724–1747. https://doi.org/10.1007/s10489-018-1332-x.

[6] Bartsch-Spörl, B., Lenz, M., & Hübner, A. (1999). Case-based reasoning: Survey and future directions. *Lecture Notes In Computer Science, 1570*, 67–89. https://doi.org/10.1007/10703016_4.

[7] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

[8] Cunningham, P. (1998). CBR: Strengths and weaknesses. In Pasqual del Pobil A., Mira J.,

& Ali M. (Eds.), *Tasks and methods in applied artificial intelligence* (pp. 517–524). Springer. https://link.springer.com/chapter/10.1007/3-540-64574-8_437#citeas.

[9] Cunningham, P. (2009). A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering, 21*(11), 1532–1543. https://doi.org/10.1109/TKDE.2008.227.

[10] De Mantaras, R. L., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Keane, M., Aamodt, A., & Watson, I. (2005). Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review, 20*(3), 215–240. https://doi.org/10.1017/S0269888906000646.

[11] Dua, D., & Graff, C. (2017). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml.

[12] Faia, R., Pinto, T., Abrishambaf, O., Fernandes, F., Vale, Z., & Corchado, J. M. (2017). Case-based reasoning with expert system and swarm intelligence to determine energy reduction in buildings energy management. *Energy and Buildings, 155*, 269–281. https://doi.org/10.1016/j.enbuild.2017.09.020.

[13] González-Briones, A., Prieto, J., Prieta, F. D. L., Herrera-Viedma, E., & Corchado, J. M. (2018). Energy optimization using a case-based reasoning strategy. *Sensors, 18*(3), 865. https://doi.org/10.3390/s18030865.

[14] Halim, Z., & Uzma, U. (2018, March). Optimizing the minimum spanning tree-based extracted clusters using evolution strategy. *Cluster Computing, 21*(1), 377–391. https://doi.org/10.1007/s10586-017-0868-6.

[15] Halim, Z., Waqas, M., Baig, A. R., & Rashid, A. (2017, November). Efficient clustering of large uncertain graphs using neighborhood information. *International Journal of Approximate Reasoning, 90*, 274–291. https://doi.org/10.1016/j.ijar.2017.07.013.

[16] Halim, Z., Waqas, M., & Hussain, S. F. (2015, October). Clustering large probabilistic graphs using multi-population evolutionary algorithm. *Information Sciences, 317*, 78–95. https://doi.org/10.1016/j.ins.2015.04.043.

[17] Haouchine, M.-K., Chebel-Morello, B., & Zerhouni, N. (2008). Competence-preserving case-deletion strategy for case-base maintenance. *European Conference on Case-Based Reasoning, 1*, 171–184. https://hal.archives-ouvertes.fr/hal-00326950/document.

[18] He, C., Tang, Y., Liu, H., Fei, X., Li, H., & Liu, S. (2019, January). A robust multi-view clustering method for community detection combining link and content information. *Physica A: Statistical Mechanics and Its Applications, 514*, 396–411. https://doi.org/10.1016/j.physa.2018.09.086.

[19] Khan, C., & Khan, M. J. (2016). Exploiting fuzzy clustering and case-based reasoning for autonomic managers. *International Conference on Autonomic Computing, Würzburg, Germany.* https://ieeexplore.ieee.org/document/7573137.

[20] Khan, M. J. (2014). Applications of case-based reasoning in software engineering: A systematic mapping study. *IET Software, 8*(6), 258–268. https://doi.org/10.1049/iet-sen.2013.0127.

[21] Khan, M. J., Awais, M. M., & Shamail, S. (2007, August). Achieving self-configuration capability in autonomic systems using case-based reasoning with a new similarity measure. *Communications in Computer and Information Science, Springer Berlin Heidelberg, 2*, 97–106. https://link.springer.com/chapter/10.1007/978-3-540-74282-1_12.

[22] Khan, M. J., Awais, M. M., & Shamail, S. (2008a, March). Enabling self-configuration in autonomic systems using case-based reasoning with improved efficiency. *International Conference on Autonomic and Autonomous Systems*, 112–117. https://ieeexplore.ieee.org/document/4488331.

[23] Khan, M. J., Awais, M. M., & Shamail, S. (2008b). Self-configuration in autonomic systems using clustered CBR approach. *International Conference on Autonomic Computing*, 211–212. https://ieeexplore.ieee.org/document/4550848.

[24] Khan, M. J., Awais, M. M., & Shamail, S. (2010). Improving efficiency of self-configurable autonomic systems using clustered CBR approach. *IEICE Transactions on Information and Systems, 93*(11), 3005–3016. https://doi.org/10.1587/transinf.E93.D.3005.

[25] Khan, M. J., Hayat, H., & Awan, I. (2019). Hybrid case-base maintenance approach for large-scale case-based reasoning systems. *Human-centric Computing and Information Sciences, 9*(1). https://doi.org/10.1186/s13673-019-0171-z.