

UNIVERSITI TEKNOLOGI MARA

**A NEW INTERNAL CLUSTER
VALIDITY INDEX FOR FUZZY C-
MEANS ALGORITHM**

KHAIRUL NURMAZIANNA BINTI ISMAIL

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science)

Faculty of Computer and Mathematical Sciences

May 2025

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student : Khairul Nurmazianna Binti Ismail

Student I.D. No. : 2019689572

Programme : Doctor of Philosophy (Computer Science) – CS950

Faculty : Computer and Mathematical Sciences

Thesis Title : A New Internal Cluster Validity Index for Fuzzy C-Means Algorithm

Signature of Student :

Date : May 2025

ABSTRACT

Fuzzy c-means is a popular clustering algorithm that can partition a set of objects into groups so that objects within one group are similar and dissimilar to others. It is required to measure cluster quality formed by any clustering algorithm, either by external or internal validations. External validation is straightforward, requiring external information such as class labels and the number of classes. Thus, external cluster validations such as Rand, F-Measure, and Fowlkes-Mallows indices serve as the ground truth in measuring cluster quality. Most external validations are based on the bounded-index model, indicating their accuracy scores from a minimum of 0 to a maximum of 1. However, in the absence of external information, e.g., unknown class and number of classes, particularly in real-world data, internal validation is then required. Instead, some internal validations are based on the unbounded-index model, which relies on a single score, whether minimum or maximum scores and no boundary. For the bounded-index models such as the Silhouette index, the Fuzzy Silhouette (FS) Index, and the Generalize Intra-Inter Silhouette (GIIS) Index, their accuracy scores range from -1 to +1, and they are different from the bounded-index model of external validation in which ranging from 0 to 1. Hence, it is difficult to benchmark any clustering algorithms or clustering results due to these constraints of the existing internal validations. In addition, no internal validation proposed thus far is associated with a model ranging from 0 to 1, like the external validation that serves as the clustering ground truth. Thus, to overcome these limitations, a new internal validation, namely, the fuzzy validity index (FVI), is proposed. The novel FVI was designed with a bounded-index model (0 to 1), ensuring interpretability and consistency with external validation indices. Beyond its bounded-index model, FVI introduces Adjusted Fuzzy Membership (AFM) and the Compactness-Separation Ratio (CSR), providing a more adaptive assessment of FCM clustering quality. These components allow FVI to effectively balance membership uncertainty and cluster separability, making it particularly suitable for fuzzy clustering applications. Seven clustering properties such as “the number of clusters”, “cluster overlap”, “dimensions and cluster overlap”, “cluster dimensions”, “cluster structure”, “unbalanced data” as well as real-world datasets (in a total of 35 datasets) were used to evaluate the performance of FVI thoroughly. Experimental results showed that the FVI is highly promising. Overall, the scores of the FVI were comparable to the scores obtained by the external validity index, i.e., F-measure. Statistically, the correlation coefficient between the FVI and F-measure was high (around 0.9 and above), indicating their similarity. Therefore, the FVI can be used as an internal validation index for FCM and could potentially serve as a ground truth for measuring cluster quality. In the future, with its strong performance and generalized formulation, FVI has the potential to be applied as an internal validation index for various fuzzy clustering techniques across different applications.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Gracious and Most Merciful, First and foremost,

I extend my deepest gratitude to my main supervisor, Dr. Ali Seman, for his steadfast guidance, invaluable advice, and continuous support throughout this research journey and the completion of my thesis. His fervent enthusiasm and dedication to research not only enlightened me in the technical realm but also imparted invaluable life lessons. His emphasis on prioritizing quality over quantity has been a guiding principle throughout. Without his meticulous mentorship, this work might not have been accomplished. I am profoundly thankful for his efforts to educate me despite my limitations.

I also extend sincere appreciation to Ts. Dr. Khyrina Airin Abu Samah, my co-supervisor, for her invaluable assistance, advice and constant moral support throughout this journey.

This work received partial support from Universiti Teknologi MARA (UiTM) and the Ministry of Higher Education Malaysia (MOHE) under the SLAB/SLAI scholarship. I am deeply grateful for their financial support, without which this research would not have been possible.

Special acknowledgments are dedicated to my beloved mother, and to the memory of my late father, Ismail bin Abdul Hamid. Thank you, Mama, for being by my side from the very beginning of this journey (the day of scholarship interview) until the end (VIVA), and for all your prayers that have continuously lifted and strengthened me.

To my dear husband, Mohd Fadli bin Saidin - thank you for being a part of this journey. The experiences and challenges throughout this period have shaped me into a more resilient and determined person, and for that, I am truly grateful.

To my beloved sons, Ammar Safwan and Haris Akhtar - may this achievement serve as a reminder that with faith, effort, and perseverance, nothing is beyond reach. I hope both of you will grow up knowing that your dreams are always worth chasing, just as I have pursued mine with all my heart.

To my siblings and family members - thank you for all the love, constant support throughout this journey.

Lastly, to my friends - whether we met at the faculty, on the field, at school, or in the postgraduate lab, each of you has played a significant part in this journey. You know who you are, and I am deeply grateful for your presence and support. A special thank you to my dear friends in the moms' circle, who have shared in both the struggles and joys of balancing family life while pursuing our PhDs together. Your support has been unwavering, and your strength has been an inspiration. For those of you still pushing forward, keep going. I'll be cheering you on at the finish line.

I am truly blessed, and I thank you Allah SWT for this and His countless blessings.

TABLE OF CONTENTS

	Page
CONFIRMATION BY PANEL OF EXAMINERS	ii
AUTHOR'S DECLARATION	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xiii
LIST OF SYMBOLS	xvi
LIST OF ABBREVIATIONS	xvii
CHAPTER 1 INTRODUCTION	1
1.1 Introduction	1
1.1.1 Fuzzy <i>c</i> -means	2
1.1.2 Cluster Validation	3
1.1.3 Clustering Ground Truth	5
1.2 Problem Statement	6
1.3 Research Objectives	8
1.4 Scopes	8
1.5 Contribution to Knowledge	8
1.6 Thesis Organisation	9
CHAPTER 2 LITERATURE REVIEW	11
2.1 Introduction	11
2.2 Cluster Analysis	11
2.3 The Partitional Clustering	12
2.4 Cluster Validations	17
2.5 External Validation	19
2.5.1 External Validation Concept	19
2.5.2 External Validation Indices	31