IRJSE International Research Journal of Science & Engineering
Int. Res. J. of Science & Engineering, Volume 14 (1) 2026
ISSN: 2322-0015
SJIF Impact Factor 6.137
https://doi.org/10.5281/zenodo.19712721
RESEARCH ARTICLE
OPEN ACCESS

Applications of Set Theory in Data Classification: A Mathematical Perspective

Arpit A. Kharwade and Rahul A. Parshuramkar*

Mahatma Gandhi Arts, Science and Late N. P. Commerce College, Armori, Gadchiroli, MS, India
*Corresponding author

Manuscript Details
Received :02.02.2026
Accepted: 26.02.2026
Published: 28.02.2026

Available online on https://www.irjse.in
ISSN: 2322-0015

Cite this article as:
Arpit A. Kharwade and Rahul A. Parshuramkar. Applications of Set Theory in Data Classification: A Mathematical Perspective, *Int. Res. Journal of Science & Engineering,* 2026, Volume 14(1): 35-41
https://doi.org/10.5281/zenodo.19712721

Abstract

Data classification, a cornerstone of modern machine learning and data mining, fundamentally addresses the challenge of assigning objects to predefined categories based on observed attributes. While contemporary approaches predominantly rely on statistical and neural methodologies, this article argues for the continuing relevance and mathematical elegance of set-theoretic frameworks in classification. We examine how foundational concepts such as subsets, partitions, equivalence relations, and complementation provide a rigorous language for conceptualizing classification problems. The article explores three principal frameworks: classical rough set theory for handling inconsistency, fuzzy rough sets for managing gradual and imprecise boundaries, and soft set theory for parameterized reasoning. We demonstrate that these approaches offer compelling advantages including interpretability, mathematical transparency, and robust handling of uncertainty. Through analysis of attribute reduction, rule induction, and granular computing, we show that set-theoretic methods provide complementary tools to mainstream algorithms, particularly when explainability and uncertainty quantification are paramount. The findings suggest that renewed attention to set-theoretic foundations can yield benefits in contemporary classification challenges involving complex, noisy, or incomplete data.

Keywords: Set Theory, Data Classification, Rough Sets, Fuzzy Sets, Granular Computing, Attribute Reduction, Rule Induction

1. Introduction

The task of data classification—assigning previously unseen instances to predefined categories based on observed characteristics—is fundamental to intelligent systems. From medical diagnosis and financial risk assessment to image recognition and spam filtering, classification algorithms underpin countless real-world applications. In recent years, the field has been dominated by sophisticated statistical models and neural architectures, achieving remarkable empirical success. However, this dominance has come at a cost: many state-of-the-art classifiers operate as opaque "black boxes," offering little insight into their decision-making processes [1].

This article revisits an alternative tradition that builds upon the mathematical foundations of set theory. The premise is straightforward yet profound: classification problems, at their core, are about partitioning a universe of objects into subsets based on shared characteristics. Set theory provides the natural language for describing such partitions, for reasoning about relationships between groups, and for establishing boundaries between categories. By embracing this set-theoretic perspective, we gain access to a rich theoretical apparatus that includes subsets, power sets, equivalence relations, complementation, and Cartesian products [2].
The relevance of set-theoretic approaches to classification extends beyond mere theoretical interest. Modern data challenges—inconsistency, incompleteness, ambiguity, and scale—demand frameworks that can gracefully handle uncertainty while maintaining interpretability. Rough set theory, introduced by Pawlak [1], offers mechanisms for dealing with inconsistent information by approximating concepts through lower and upper approximations. Fuzzy set theory, developed by Zadeh [9], captures the gradual nature of category membership. The synthesis of these traditions in fuzzy rough set theory provides powerful tools for managing the multifaceted uncertainties present in real-world data [3]. More recent developments, such as soft set theory and granular-ball computing, extend these foundations further [4, 8].

This article aims to provide a comprehensive mathematical exploration of set-theoretic applications in data classification. We shall examine how fundamental set operations translate into classification operations, how rough sets handle inconsistency, how fuzzy sets manage vagueness, and how these frameworks can be combined to create robust classifiers. The article does not claim that set-theoretic methods should replace existing approaches but rather that they offer valuable complementary capabilities, particularly when interpretability, mathematical transparency, and uncertainty handling are essential.

The structure of the article is as follows: Section 2 establishes the foundational set-theoretic concepts relevant to classification. Section 3 introduces rough set theory and its application to classification with inconsistent data. Section 4 extends the discussion to fuzzy and fuzzy rough approaches for handling gradual boundaries. Section 5 examines soft set theory and its role in rule-based classification. Section 6 explores set-theoretic perspectives on feature selection and dimensionality reduction. Section 7 addresses the challenges and limitations of these approaches, while Section 8 presents concluding remarks and directions for future research.

2. Foundational Set-Theoretic Concepts in Classification

2.1 The Classification Problem in Set-Theoretic Terms

A classification problem can be formally described using the language of sets. We begin with a universe of discourse, U, representing all possible objects under consideration. Each object $x \in U$ is characterized by a set of attributes $A_1, A_2, \ldots, A_n$, with each attribute A_j taking values from its domain $dom(A_j)$. An object is thus represented as a vector $a^i=(a_1^i, a_2^i, \ldots, a_n^i)$, where a_j^i denotes the value of attribute A_j for object i [2].

The set of classes $C=\{C_1, C_2, \ldots, C_k\}$ forms a partition of U, meaning that every object belongs to exactly one class and each class is non-empty. Formally:
$U = \bigcup_{(m=1)}^{k} C_m$, $C_m \cap C_l = \emptyset$ for $m \neq l$

The classification task is to construct a classifier $f: U \to C$ that, given the attribute values of an object, assigns it to the correct class. In practice, however, we do not have access to the true classification function; instead, we observe a finite training set $T \subseteq U$ with known class labels, and we must induce a classifier that generalizes to unseen objects [2].

2.2 Set Operations and Their Classification Analogues
Several fundamental set operations have direct analogues in classification:

Subset and Superset Relations: The notion that one class is contained within another corresponds to hierarchical classification. For instance, if we define "animal" as a superset of "mammal," then any object classified as a mammal should also be classifiable as an animal.

Partitions and Equivalence Relations: The classification of objects into mutually exclusive categories is exactly a partition of the universe. Equivalence relations—reflexive, symmetric, and transitive relations—provide the mathematical basis for partitions. Objects related by an equivalence relation share the same class label [2].

Complementation: For any class $C \subseteq U$, its complement $U \setminus C$ represents all objects not belonging to that class. In binary classification, the complement is precisely the negative class. This simple observation underpins many classification algorithms: to distinguish between class C and its complement, we seek attributes that separate the two subsets [2].

Cartesian Products: The description space $A = \mathrm{dom}(A_1) \times \mathrm{dom}(A_2) \times \cdots \times \mathrm{dom}(A_n)$ is a Cartesian product of attribute domains. Classification rules can be viewed as subsets of this description space that map to class labels.

2.3 The Logic of Classification Rules
A fundamental insight is that classification rules can be expressed as logical statements using set-theoretic notation. For a class C with examples E and counter-examples E', a concept C that is consistent with the data must satisfy [2]:

$$E \subseteq C \subseteq \overline{(E')}$$

where $\overline{(E')}$ denotes the complement of the counter-examples. This elegant formulation captures the essence of concept learning: the class C must include all observed examples and exclude all observed counter-examples. The lower bound E represents the most specific concept consistent with the data, while the upper bound $\overline{(E')}$ represents the most general concept.

In logical terms, this becomes [2]:

$$\bigvee_{(i:e^{i}\in E)} \left(a_1^{i} \wedge a_2^{i} \wedge \cdots \wedge a_n^{i} \right) \vDash v\_C \vDash \bigwedge_{(j:e'^{j}\in E')} \left(\neg a_1'^{j} \vee \neg a_2'^{j} \vee \cdots \vee \neg a_n'^{j} \right)$$

where v_C is the propositional variable indicating membership in class C. This logical formulation demonstrates that the set-theoretic approach provides both a combinatorial and logical foundation for classification.

3. Rough Set Theory: Handling Inconsistency

3.1 The Challenge of Inconsistent Information
Real-world data rarely conform to simple, deterministic classification rules. Inconsistency arises when objects with identical attribute values belong to different classes. Such situations are inevitable in practice, whether due to measurement error, incomplete information, or inherent ambiguity in classification.
Set theory offers a principled approach to handling inconsistency through the concept of rough sets. A rough set is an approximation of a concept using two sets: a lower approximation (the set of objects definitely belonging to the concept) and an upper approximation (the set of objects possibly belonging to the concept) [1].

3.2 Formal Foundations of Rough Sets
Let U be a universe of objects and R be an equivalence relation on U representing the indiscernibility of objects based on their attributes. The equivalence classes of R partition U into subsets of objects that are indistinguishable given the available attributes. These equivalence classes are the fundamental granules of information in the data [1].
For a target concept $X \subseteq U$, the lower approximation $\underline{R}X$ and upper approximation $\overline{R}X$ are defined as [1]:

$$\underline{R}X = \{x \in U : [x]R \subseteq X\}$$
$$\overline{R}X = \{x \in U : [x]R \cap X \neq \emptyset\}$$

where $[x]R$ denotes the equivalence class of x under relation R.

The lower approximation consists of equivalence classes fully contained in X, representing objects that can be confidently assigned to the concept. The upper approximation consists of equivalence classes that intersect X, representing objects that may belong to the concept. The boundary region Bnd(X) = $\overline{R}X \setminus \underline{R}X$ contains objects that cannot be definitively classified based on the available information [1].

3.3 Rough Set Approaches to Classification

In classification contexts, rough sets offer multiple benefits:

Handling Inconsistent Data: When objects with identical attribute descriptions belong to different classes, the equivalence classes cross class boundaries. Rough sets represent this ambiguity explicitly through the boundary region, allowing classifiers to handle inconsistency without forcing arbitrary decisions [1].

Rule Extraction: Rough sets facilitate the extraction of deterministic and approximate rules. From the lower approximation, we can derive certain rules that hold with certainty. From the boundary region, we can derive possible rules that hold with some degree of uncertainty [1].

Attribute Reduction: A critical contribution of rough set theory is the concept of reducts—minimal subsets of attributes that preserve the classification power of the full attribute set. This addresses the dimensionality reduction challenge that is essential for building efficient classifiers [6].

3.4 Granular-Ball Rough Sets: A Recent Advancement

Recent research has extended rough set theory through the concept of granular-ball computing. In this framework, data is represented as "granular-balls"—clusters of similar instances—rather than individual points. The granular-ball fuzzy rough set model defines lower and upper approximations using these granular-balls, which enhances robustness against noise and provides smoother classification boundaries [4].

The granular-ball approach addresses limitations of traditional fuzzy rough sets, including limited granularity control, sensitivity to noise, and difficulties handling large-scale datasets. By grouping similar instances together, granular-ball computing offers improved interpretability and more adaptive handling of uncertainty [4].

4. Fuzzy Sets and Fuzzy Rough Sets

4.1 Fuzzy Sets: Handling Gradual Membership

While classical set theory assumes crisp boundaries between categories, real-world categories often have gradual membership. A patient may be "somewhat" at risk for a disease, or a financial transaction may be "partially" suspicious. Fuzzy set theory addresses these situations by allowing membership degrees in the interval [0, 1] rather than binary values of 0 or 1 [9].
Formally, a fuzzy set A in a universe U is characterized by a membership function $\mu\_A: U \rightarrow [0,1]$, where $\mu\_A(x)$ represents the degree to which element x belongs to A [9].

In classification, fuzzy sets enable us to model classes with gradual boundaries. An object can belong to multiple classes to different degrees, reflecting the ambiguity inherent in many classification tasks [9].

4.2 Fuzzy Rough Sets: Synergistic Uncertainty Handling

Fuzzy rough set theory integrates the imprecision-handling capabilities of fuzzy sets with the inconsistency-handling capabilities of rough sets [3]. The resulting framework provides a comprehensive approach to managing both types of uncertainty in data.
In fuzzy rough sets, the equivalence relations of classical rough sets are replaced by fuzzy similarity relations. The lower and upper approximations become fuzzy sets, with membership degrees indicating the certainty with which an object belongs to a concept [3].
The mathematical formulation extends the rough set definitions:

For a fuzzy set A and a fuzzy similarity relation R, the fuzzy lower and upper approximations are [3]:

$$\underline{R}A(x) = \inf\_(y\in U)\ \{I(R(x,y),\ A(y))\}$$
$$\overline{R}A(x) = \sup\_(y\in U)\ \{T(R(x,y),\ A(y))\}$$

where I is a fuzzy implication operator and T is a t-norm [3].

4.3 Fuzzy Rough Neural Networks

Recent research has sought to integrate fuzzy rough set theory with neural networks, creating adaptive frameworks that combine mathematical rigor with learning capabilities [7]. The fuzzy rough neural network model adaptively learns fuzzy similarity relations through backpropagation, addressing the lack of adaptive learning in traditional fuzzy rough sets [7].

The network architecture includes layers for [7]:

- Input representation
- Membership computation
- Fuzzy lower approximation
- Feature weighting
- Output classification

This integration demonstrates the synergies possible between set-theoretic and neural approaches, suggesting that these traditions need not be in conflict [7].

5. Soft Set Theory in Rule-Based Classification

5.1 Fundamentals of Soft Sets

Soft set theory, introduced by Molodtsov, offers a parameterized approach to set-theoretic reasoning [8]. A soft set is a pair (F, P), where P is a set of parameters and F: P $\rightarrow$ P(U) is a mapping from parameters to subsets of the universe [8].

The parameterized nature of soft sets makes them particularly suitable for representing multi-valued information systems, where different attributes (parameters) provide different perspectives on classification. Unlike rough sets, which rely on equivalence relations, soft sets are more flexible in handling diverse types of information [8].

5.2 Multi-Soft Sets and Classification

Multi-soft sets extend the concept to handle multiple-valued information systems. In this framework, each attribute can be viewed as defining a soft set, and the combination of soft sets provides a comprehensive representation of the data [8].

For rule-based classification, multi-soft sets offer an intuitive approach [8]:

- Each attribute-value combination defines a subset of objects
- Classification rules are formed by combining these subsets through set operations
- The classifier is constructed by identifying rules that accurately separate classes

5.3 Integrating Soft Sets with Decision Trees

Soft set theory has been combined with decision tree induction to create transparent classifiers [8]. The soft set representation provides a natural bridge between the training data and the decision tree structure, where nodes represent tests on parameters and branches represent outcomes [8].

This integration offers several advantages [8]:

- Parameterized representation accommodates complex attribute structures
- Set operations provide a mathematical foundation for rule extraction
- The resulting classifiers maintain interpretability while handling complex data

6. Attribute Reduction and Feature Selection

6.1 The Set-Theoretic Perspective on Dimensionality

Dimensionality reduction is essential for building efficient classifiers from high-dimensional data. From a set-theoretic perspective, the goal is to find minimal subsets of attributes that preserve the classification structure of the data [6].

The concept of a reduct is central to this perspective. Given an information system with conditional attributes and decision attributes, a reduct is a subset of conditional attributes that has the same classification power as the full attribute set [6].

6.2 Rough Set-Based Attribute Reduction

Rough set theory provides a principled approach to finding reducts. The dependency degree $\gamma\_P(D)$

measures how well the decision attribute D can be predicted from the conditional attributes P [6]:

$$\gamma\_P(D) = (|POS\_P(D)|) / (|U|)$$

where POS_P(D) is the positive region—the set of objects that can be definitively classified using attributes in P [6].

A reduct is a subset R⊆P that preserves dependency:

$$\gamma\_R(D) = \gamma\_P(D) \text{ [6].}$$

Finding minimal reducts is NP-hard, motivating heuristic approaches. One approach uses relational algebra operations—projection and division—to compute scores for attributes based on their information content. Attributes are partitioned into groups according to these scores, and reducts are generated through iterative group processing [6].

6.3 Rule Generation and Pruning

Once reducts have been identified, classification rules can be generated. Each reduct provides a basis for rule extraction: the rules are combinations of conditional attribute values that imply a class label [6].

The rule generation process typically involves [6]:

- Applying a decision tree algorithm to each reduct
- Extracting rules from the resulting trees
- Pruning rules by removing extraneous components
- Selecting a minimal set of rules that achieves good classification accuracy

The set-theoretic perspective ensures that the rules are mathematically sound and can be validated against the training data [6].

7. Challenges and Future Directions

7.1 Challenges of Set-Theoretic Approaches

Despite their theoretical elegance, set-theoretic approaches to classification face several challenges:

Scalability: Computing reducts and equivalence classes can be computationally expensive for large datasets. This is a significant limitation in the era of big data [10].

Sensitivity to Discretization: Rough set approaches require discrete data; continuous attributes must be discretized, which can lead to information loss [1].

Noise Sensitivity: Pure set-theoretic approaches can be sensitive to noise in the data. A single mislabeled example can dramatically affect the boundary regions [4].

Limited Adaptability: Traditional rough and fuzzy rough sets lack adaptive learning capabilities, though recent work on fuzzy rough neural networks is addressing this limitation [7].

7.2 Emerging Directions

Several promising directions are emerging:

Integration with Machine Learning: The development of fuzzy rough neural networks and similar hybrid approaches suggests that set-theoretic methods can complement rather than compete with mainstream machine learning [7].

Granular Computing: The granular-ball framework represents a significant advance in making rough set approaches more robust and scalable [4].

Dynamic and Incremental Learning: Recent work on dynamic fuzzy-rough sets addresses the challenge of large datasets through incremental learning mechanisms [10].

Set-Valued Classification: The recognition that classifiers need not output a single label but can return sets of possible classes provides another direction where set theory naturally contributes [11].

8. Conclusion

This article has explored the mathematical foundations of set-theoretic approaches to data classification. We have examined how fundamental set operations—subsets, partitions, equivalence relations, and complementation—provide a rigorous language for conceptualizing classification problems [2]. The frameworks of rough sets [1], fuzzy sets [9], and soft sets [8] each offer distinct capabilities for handling the uncertainties that inevitably arise in real-world data.

Rough set theory provides a principled approach to inconsistency, representing concepts through lower and upper approximations [1]. Fuzzy set theory addresses gradual boundaries through membership degrees [9]. Their synthesis in fuzzy rough sets combines these capabilities [3], while soft set theory offers a parameterized alternative for multi-valued information systems [8]. Recent developments in granular-ball computing [4] and fuzzy rough neural networks [7] demonstrate the continued evolution of these frameworks.

The set-theoretic perspective on classification offers several compelling advantages: mathematical transparency, rigorous handling of uncertainty, and the capability to extract interpretable rules. These qualities are increasingly valuable as applications demand both accuracy and explainability. However, challenges remain in scalability [10], noise handling [4], and integration with contemporary machine learning pipelines.

We conclude that set-theoretic methods should be viewed not as competitors to mainstream classification approaches but as complementary tools. The mathematical rigor of set theory provides a foundation upon which hybrid approaches can be built, combining the strengths of different paradigms [7]. As the field moves toward more interpretable and uncertainty-aware AI, set-theoretic frameworks will likely play an increasingly important role [11].

Conflicts of interest: The authors stated that no conflicts of interest.

Correspondence and requests for materials should be addressed to **Rahul A. Parshuramkar**

Peer review information : IRJSE thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at https://www.irjse.in/reprints

References

1. Pawlak Z. Rough sets. International Journal of Computer & Information Sciences. 1982;11:341-356.
2. Dubois D, Prade H. Towards a logic-based view of some approaches to classification tasks. In: Advances in Intelligent Data Analysis. 2020:1-16.
3. Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems. 1990;17:191-209.
4. Xia S, et al. On granular-ball fuzzy rough sets and applications in attribute evaluations. Fuzzy Sets and Systems. 2025;519.
5. Chaudhuri A. Fuzzy rough support vector machine for data classification. International Journal of Fuzzy System Applications. 2016;5(2):26-53.
6. Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Intelligent Decision Support. Springer; 1992:331-362.
7. Zhang Q, et al. An adaptive fuzzy rough neural network and its application in classification. IEEE Transactions on Fuzzy Systems. 2025.
8. Singh PK. Soft set theory in rule-based classification. Ph.D. dissertation. 2021.
9. Zadeh L. Fuzzy sets. Information and Control. 1965;8:338-353.
10. Zhang Y, et al. Dynamic fuzzy-rough sets for uncertainty-aware knowledge extraction from large datasets. In: Proc. IEEE International Conference on Data Mining. 2025.
11. Karlsson M, Hössjer O. Classification under partial reject options. Journal of Classification. 2024;41:2-37.

Publisher's Note

IJLSCI remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.