ROUGH SETS, SIMILARITY, AND OPTIMAL APPROXIMATIONS

ROUGH SETS, SIMILARITY, AND OPTIMAL APPROXIMATIONS

BY

ADAM LENARČIČ,

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTING & SOFTWARE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2017) McMaster University

(Computing and Software) Hamilton, Ontario, Canada

TITLE:	Rough Sets, Similarity, and Optimal Approximations
AUTHOR:	Adam Lenarčič
	B.Sc., (Computer Science and Mathematics)
	Brock University, St.Catharines, Canada
	M.Sc., (Computer Science)
	Brock University, St.Catharines, Canada
SUPERVISOR:	Dr. Janicki
NUMBER OF PAGES:	vi, 102

Abstract

Rough sets have been studied for over 30 years, and the basic concepts of lower and upper approximations have been analysed in detail, yet nowhere has the idea of an 'optimal' rough approximation been proposed or investigated. In this thesis, several concepts are used in proposing a generalized definition: measures, rough sets, similarity, and approximation are each surveyed. Measure Theory allows us to generalize the definition of the 'size' for a set. Rough set theory is the foundation that we use to define the term 'optimal' and what constitutes an 'optimal rough set'. Similarity indexes are used to compare two sets, and determine how alike or different they are. These sets can be rough or exact. We use similarity indexes to compare sets to intermediate approximations, and isolate the optimal rough sets. The historical roots of these concepts are explored, and the foundations are formally defined. A definition of an optimal rough set is proposed, as well as a simple algorithm to find it. Properties of optimal approximations such as minimum, maximum, and symmetry, are explored, and examples are provided to demonstrate algebraic properties and illustrate the mechanics of the algorithm.

Contents

Acknowledgements

First and foremost I would like to thank my supervisor, Dr. Ryszard Janicki. If he did not accept my PhD proposal, and then guide me through this process, none of this would have been possible. Next, I thank Dr. Michael Soltys and Dr. Ridha Khedri for their instruction while I researched areas of interest, and for their support when I proposed this thesis plan. I also acknowledge and thank my entire supervisory committee for their insight and suggestions during the revision process. My friends Andrew and Troy deserve credit for improving my grammar and overall clarity. I must also mention the support of my partner, Hannia, who has continually encouraged my work on this thesis. Lastly, I thank my parents for their emotional and financial support without which I would not have had the freedom and courage to pursue this PhD. Without the continuous encouragement and support of everyone, this thesis would not be as well written as I believe it to be.

Chapter 1

Introduction

Approximation is inherent in the capture of all data. If reference is made to anything at all, the reference itself must be approximated in some way. Surely every detail of an environment, and its context, cannot be *fully* captured, owing to limitations of representation and storage.

If someone tried to specify exactly every detail of a situation, object, or concept, a cunning devil's advocate could surely list the ways in which that specification was insufficient, inaccurate, or imprecise. For example, while a photograph does provide some information about a scenario, it lacks multiple points of view, and so approximates a scene from only one perspective. Due to optical illusions and camera tricks we cannot be truly certain how far each item is to the single point of view. Conversely, a range finder could plot distance from itself to every point in the scene, record the colour and simulate an environment precisely. But this is still only from one point of view, meaning we have to deal with occlusion. So it could then be moved around to provide multiple perspectives, but inevitably some part of a scene is not recorded, nor are the contents (insides) of objects observed, so things like density are still not

known. Robotic perception and representation at its most fundamental levels are essentially approximation. Some level of truncation is involved in assigning value to the continuous reality around us as machines have only limited storage capacity. Even with what we *can* capture, the data recorded is inevitably only precise to a certain number of digits stored in the electronic devices, and there is always a degree of precision and accuracy associated with the detection and recording mechanisms.

Sometimes, we don't even care about precision which could easily be specified. If someone asks what time it is, your answer is surely rounded to some relevant digits. Rarely do the number of seconds matter in our daily lives, so we approximate our answer by rounding to the nearest fifteen or half hour. If we're waiting for the bus or train, we might care about the number of minutes. Exact time is rarely specified, because the smallest units are so fleeting, so it is approximated by hours and possibly minutes, yet in certain instances such as a competition setting, or the tracking of satellites, both precision and accuracy are vital to tiny fractions of milliseconds. We tend to approximate based on the granularity of the frame of reference. When numerically representing the real world, the amount of precision and the size of storage are always two opposing goals to be optimized, and since storage cannot be infinite, we must always admit some non-zero compromise in precision and thus the representation is necessarily an approximation of the true value.

In other cases, categories defined by attributes are given and the task is to find which category or set of categories fits best. If one were to empty a filled garage or storage unit they could start sorting things by colour, or size, but usually people separate by category. Put tools together, liquids together, things for the car, holiday decorations, etc. In a hospital, patients are separated by type of care they need, or

by their symptoms, not by their attributes such as height, weight, or hair colour. By grouping patients by common attributes, classes or granules are created as basic building blocks. If we need to isolate which group of patients are most susceptible to a newly occurring disease, we can use the known people diagnosed to form a target set, and we can find the class of people based on symptoms or other gathered attributes who are at highest risk to contract the disease. Of course, refinements would be necessary, but it would prove invaluable to narrow an extremely large list to a class that is a fraction of the size. With enough attributes it could even be done iteratively. The ability to isolate the best set of classes of equivalent attributes to approximate a target set is easily seen to be a valuable tool in any machine learning arsenal.

If a problem is continuous and lacks obvious classes, it can be analyzed by discretizing the problem using different category sizes which could be refined by testing several sizes, and comparing which gives the largest optimal similarity. In any case, this maximum similarity for a particular equivalence class and type of measurement is referred to as the *Optimal Approximation*.

In what follows, it is shown how to efficiently determine the best group of classes to approximate a particular target set.

1.1 Motivation and Rationale

Since Pawlak published the first papers on rough sets, the concepts of upper and lower approximations have been fundamental to all works in the rough set context[21]. Given a continuous figure but discrete elements, then the approximation process involves choosing which discrete elements should be included in devising the best approximation. Those which we are sure will be included, since they satisfy all

criteria, fall in the upper approximation. While those which we are unsure about make up the border or boundary region. The concepts of strong and weak membership naturally correspond to upper and lower approximations respectively. For more detail see Section 3.2.1.

One obvious application was to group data, and at times there are often concepts which can not be defined rigidly. There is still a need to express this using the categories alone. The idea is to form a group of all of the possible ways the concept might look based on the data which was available. So in its most basic sense, a lower approximation is a representation of all the information we absolutely know are true about something, and an upper approximation is a representation of all the information that might possibly be true. What has lacked until now has been any way of expressing the group of building blocks partitioned by the rough sets which is most representative of a non-rigid set. This thesis seeks to introduce a way to express and find this "best rough set approximation."

The introduction of an algorithm to find an optimal rough approximation to a non-exact target set will allow researchers in the field of rough sets to test for and compute the 'best' comparison instead of the current possibilities which only include whether a set is contained in an upper/lower approximation, or if sets are roughly equal. Nowhere in the rough set literature has it been proposed to isolate the 'most similar' rough set to a given non-rough set, so this thesis not only developed a definition, but also an algorithm to find it. The algorithm is actually quite fast in that it runs in linear time, except that the sets must be sorted first so due to that requirement, the algorithm proposed here will run in $O(nlogn)$ (or optimal sorting) time.

1.2 Contributions

The main contributions of this thesis are as follows:

1. A review of foundations of rough sets, similarity indexes, and measure theory.
2. Properties of comparison indexes are provided.
3. Specification of Measure Theory to require the null-free property.
4. Introduction of concepts called *Borders* (analogous to Pawlak's Boundary Region) and *Border Sets*.
5. Definition of the concept of an *optimal rough set* approximation, an extension of Rough Sets.
6. Definition of an algorithm to find the optimal rough set among a given universe and prove correctness of this algorithm.

1.3 List of Publications

A list of publications which arose from this work:

1. Optimal Approximations with Rough Sets[13]
2. Optimal Approximations with Rough Sets and Similarities in Measure Spaces[14]

1.4 Outline of Thesis

In what follows the organization of this thesis is outlined. This chapter, introduces the ideas behind this work. Chapter 2 is designed to be background reading, a sort of

foundation from which the ideas of this work branched off. A literature review of the most prominent papers, and other relevant works are provided. Chapter 3 contains mathematical background which is required to proceed with our works. Definitions from Measure Theory are provided in Section 3.1 along with a specification of Measure Theory to require the null-free property. In section 3.2 the foundations of rough sets are outlined, as well as several similarity indexes.

Chapter 4 and 5 contain the main contributions of this thesis. Chapter 4 examines axioms of similarity, and presents all the similarity indexes that are used in this thesis. Chapter 5 presents a definition of an optimal approximation within the foundation of rough sets, as well as an algorithm to find it. It also contains a brief discussion on asymmetry and the assumption that must be made about it to work in the context of rough sets, which is picked up from the literature review. Within Chapter 5 there is also a section discussing applications.

Chapter 2

Literature Review

2.1 Measures of the Amount of Ecologic Association between Species - Dice (1945)

When Lee Dice [5] noticed that no method had come into general use by ecologists to quantitatively express how "associated" two species were to one another, he took it upon himself to provide a method he judged to be superior to the alternatives. He cited several proposals, but they were either unreliable, too complex, or impractical. None were sufficient to express what he wished.

In ecology, sometimes multiple samples of an environment are taken in a standardized way to determine how often species occur together. Each species is present in a certain number of samples. When two species are being compared (i.e. their degree of association is being measured), the number of samples in which each species occurs (usually termed a and b), and the number of samples in which they occur together (h) is recorded. Along with the total number of samples (n), this yields four numbers,

and two more which represent the number of occasions where each species occurs alone may be derived by subtracting the number where they occur together from the total numbers: $a - h$ and $b - h$.

Though Dice did not express these numbers in set-theoretical terms, we could translate the representation Dice used into set theoretical terms. The total number of samples, would be the size of the universe, i.e. $n \to |U|$. The locations in which each species is detected would correspond to each set (items which are equivalent by a 'same species' relation), (e.g. A, B). The variables used by Dice for the number of samples where each species occurs, would then correspond to the magnitude/size of each set, i.e. $a \to |A|, b \to |B|$, and the number where species occur together fits nicely with the size of the intersection of the sets, i.e. $h \to |A \cap B|$. We can also see that the size of each set difference equates to the frequency with which each species occurs alone: $|A \setminus B| \to a - h, |B \setminus A| \to b - h$.

In devising his measure, Dice began with a *coefficient of association* as developed by Forbes in 1907, given by the equation $\frac{hn}{ab}$. This was based on the statistical definition of "expectation". This, unfortunately for Dice, incorporated an element of probability as it was meant to express how much more or less often the species occur together than would be expected by random chance. It is based on the assumption of uniform distribution, so if this is shown invalid all calculated values are invalid as well. Since reality is rarely uniform this assumption creates an oversimplification. If there are n samples, and species A appeared in a of the samples, the chance of species A to occur in any particular sample is a/n, and likewise with b/n. Probability theory dictates that (given independence) the probability two events occur together is the product of the probabilities each would be expected to occur, so the chance

they occur together is $a/n \times b/n = ab/n^2$. If we multiply this by the total number of samples, the expected number of samples where the species occur together is ab/n. The coefficient of association divides the actual number of recorded samples where the species occur together by this number, i.e. $\frac{h}{ab/n} = \frac{hn}{ab}$. One positive was that it could give an easily comprehensible indication of how close to expectation the species occurred together, as values smaller than 1.0 show the species occur together less than expected and values greater than 1.0 show they occurred together more than expected.

The value of this measure, however, depends on the abundance of each species in the sampled environment. As Dice noted, when both species are abundant, "both will appear in a high proportion of the samples taken. The chance of both species occurring together in any given sample is therefore high, and it will be impossible for any great deviation from expectation to occur." ([5],P.298) A similar situation occurs when both species are sparsely present in the samples. If this factor must be considered, the coefficient of association is then not general enough for use in all situations. He quoted this coefficient as, "...only a measure of the amount of deviation of the number of occurrences together from the number expected by chance." ([5],P.298)

So Dice next proposed the *association index*, which simply divided the number of times a species occurred alone by the number of times two species occurred together. He did note here that the association index,

> "may differ depending on which species is used as the basis of comparison. Such a difference may point out an important ecologic relation between the two species. Frequently one species is dependent upon another, without there being any reciprocal dependency." ([5],P.298)

This notion is also discussed in Section 5.3, since unlike this index, we deemed it necessary for our purposes to require that our similarity function be symmetrical.

Though not by name, the symmetry requirement was noted by Dice, as he reported that a measure which did not depend on a chosen base species, was needed for some ecologic studies. He then refined the above by taking "a value intermediate between the reciprocal association indexes A/B and B/A." Given by the formula $\frac{2h}{a+b}$, it was termed *the coincidence index* which has become known as the Dice, Sørensen, or Dice-Sørensen index [4]. So the overskewing of commonality (compared to the alternative Jaccard Index which would look like $\frac{h}{a+b-h}$) was not in fact purposeful. It seems Dice did not consider using the average of A/B and B/A, possibly due to its complexity ($\frac{ah+bh}{2ab}$ in its simplest form), or he could have simply overlooked the option. Yet in modern scenarios when more focus is to be placed on common elements, this index can be chosen intentionally.

Dice went on to note that values range from 0.0 to 1.0 which allows for easy comprehension. He also mentioned that while there was no statistical test of the reliability of the measures themselves, the chi-square test would indicate for any given sample, if the results might be due to random chance. To perform this test, the expected number of samples for each species alone is required (earlier we looked at all samples, included those with both species). This could be calculated by subtracting the expected number of samples where species occur together from the actual number of samples in which each species was recorded, i.e. $a-ab/n$, $b-ab/n$. Chi-square also requires the probabilities that each species will not occur in any particular sample, given by $(n-a)/n$ and $(n-b)/n$, and the chance that neither species would be found, given by $(n-a)/n \times (n-b)/n$. The expected number of samples where neither species

is detected is then $(n-a)(n-b)/n$. For each of the four groups: A only, B only, A + B, and neither, we divide the square of the difference between observed and expected values by the expected value, i.e. $\frac{deviation^2}{expectated}$, and then we sum these 4 values. The result must be compared to a Chi-square table using a degree of freedom of one (because we used data from two species), and this will yield the probability the results we obtained were due to random chance.

Dice concluded with the following comments on the reliability and usability of these measures. The magnitude of the Chi-square test suffers from the same drawback as the coefficient of association, in that it varies based on the scarcity or abundance of the species being compared, so further interpretation must be done. The association between three or more species can be determined by expansion of the formulas used for two species, but it is difficult to obtain a sample size needed. To be reliable, sample sizes for association between two species should be greater than 100; the number of expected units in each of the four classes should exceed ten and values less than five are unreliable. Other possible issues Dice listed, involve the omission how many individuals of each species occur in each sample, or the tendency of some individuals to be associated with larger groups such as families, packs, or swarms. The way samples are drawn is also of relevance, as it is important to ensure that each sample's size and scope is appropriate, and the method of collection is fair to both species.

It is also valuable to note that a high association index, coefficient index, or coefficient of association value, do not necessarily explain or define an association between species. Two species may be mutually attracted to – or dependent on – another species or other factor in their environment. Thus, in some cases associated

species may have no direct relationship beyond selection of the same habitat, while in others, some more fundamental association may be found between the species such as a predator-prey link where one animal feeds on another (even if there might be another species between them), or some sort of symbiosis where both species benefit from the others presence. □

2.2 Halmos - Measure Theory (1950)

Generally when the size of a set is judged, it is the number of elements within it (cardinality) that is measured. This, however, is only one of the ways the size of a set of items can be measured. We chose to review and use ideas presented in a graduate textbook in Mathematics called *Measure Theory*[9], written by Paul Halmos in 1950 due to its generalized treatment of measure spaces, its extremely high citation count, and the fact that much of the material remains relevant today. The book was also reprinted several times by the *Graduate Textbooks in Mathematics* series demonstrating its value in the field [10]. The book examined many aspects of generalizing measurement, but this review focuses on measure functions and spaces, which are the most relevant parts of the text.

Within the prerequisite reading, he defined a metric space as a set X and a real valued distance function d on $X \times X$ with the following three basic requirements. The distance between two elements can be zero, only if they are the same element ($d(x, y) = 0 \iff x = y$). The distance from one object to another, is the same as the reverse distance, i.e. d is reflexive ($d(x, y) = d(y, x)$). The triangle inequality is used to require that the distance between any two points is shorter than the sum of the distances through an intermediate point.

The first chapter of *Measure Theory* provided a background to familiarize the reader with basic principles in sets and classes then chapter two was dedicated to measures on structures such as rings, intervals, and of course sets. We did not use material from chapter three which focuses on extensions of measures, but chapter four outlined measure spaces, and measurable functions which play an important role in this thesis.

In our section on measure theory (Section 3.1), we define a measure space as a triple consisting of a universe, the universe's power set, and a function/measure satisfying three requirements, but this is a slight simplification. In Halmos definition, like ours, a measure space triple also begins with a set and ends with a function. Instead of the power set of the universe used here, Halmos used a more general σ-ring S of subsets of the universe, requiring that $\bigcup S = X$, i.e. every point is contained in some set, or no point exists that does not belong to any set. We do not need this requirement since we only need to be able to compare any element in the universe to any other, so we use the power set which trivially satisfies Halmos extra closure requirement. Using a σ-ring instead would allow for more information to be expressed when extra data is provided which partitions the universe, and is closed under union and complementation. For us to use the more general σ-ring structure would mean-among other things-that we would need 2 operations on the universe of elements U (call them $*$ and $\circ$) where $(U, *)$ forms an abelian group, and $(U, \circ)$ forms a monoid and $\circ$ distributes over $*$. We would need to show the existence of $*$ and $\circ$ and prove or assume that $*$ is associative, commutative, and has a zero element and $\circ$ is associative, and has an identity element. However, we wished for our work to be applicable in a situations where these binary operations are not defined or do not satisfy the axioms.

Essentially we want to be able to compare and judge the similarity of two sets of items within a larger universe, regardless of the properties, or lack thereof.

This thesis uses measure theory to create a way of assigning values to each set of items. It is then possible to define the function in a measure space to evaluate any attributes of the elements in our universe. Most often, this thesis assigns the measure to be cardinality, so that number of items in a set can be counted, but at times, it may be needed to evaluate similarity based on other attributes such as size, weight, price, etc. □

2.3 Marczewski - On a Certain Distance of Sets and Corresponding Distance of Functions (1958)

This short paper by Marczewski and Steinhaus[18] was the inspiration for the author's idea of using a foundation of measure theory as a generalization technique. The purpose of the paper was to study and report properties of the 'distance between sets' formula written below, and the corresponding 'distance between functions'. Drawing from both measure theory and distance between sets, the authors propose a new metric they refer to as σ_μ (where $\mu : 2^U \to [0, 1]$ represents a function returning the measure of the set) which is defined as the ratio of measures like so:

$$\sigma_\mu(A, B) = \begin{cases} \frac{\mu((A \setminus B) \cup (B \setminus A))}{\mu(A \cup B)} & \text{if } \mu(A \cup B) > 0 \\ 0 & \text{if } \mu(A \cup B) = 0 \end{cases}$$

They then prove that the triangle inequality holds, and explain that the maximum and minimum values of 1 and 0 respectively, follow directly from the definition. The

maximum value occurs when the two sets have an empty intersection, since then the symmetric difference and the union are equal. The minimum value is piece-wise defined to be zero whenever the measure of the union is zero, to prevent an undefined value (i.e. denominator zero). This definition of distance is equivalent to the Jaccard similarity index if we subtract the result from 1.

The example which the authors provided at the end of their paper demonstrates the contrast between evaluation using sets and evaluation using functions. This suggested that it is not possible to accurately and completely define what it means for one set to 'best' approximate another without somehow referring to which attributes are used to define the size of a set. Essentially, how the set is to be measured must be known, before it is possible to evaluate how similar/different two sets are with respect to that measurement. This is the purpose of the μ function which is defined in section 3.1 □

2.4 Zadeh - Fuzzy Sets (1965)

In 1965 Zadeh produced a foundational article with merely three citations[39]. Referencing only Lattice Theory by Birkhoff, Naive Set Theory by Halmos, and Introduction to Metamathematics by Kleene, L. A. Zadeh characterized fuzzy sets as a class of objects, each with a continuum of grade of membership. "Essentially," Zadeh wrote, "such a framework provides a natural way of dealing with problems in which the source of imprecision is the absence of sharply defined criteria of class membership rather than the presence of random variables."

Formally, given a space of objects X, a fuzzy set A is defined by a membership function $f_A(x)$ which maps each point x in X to a real number in the interval $[0, 1]$.

If A is an ordinary set, the function can only map to 0 or 1 according to whether x belongs to A.

Important concepts outlined by Zadeh include complement, containment, union, and intersection. The complement of fuzzy set A is simply $f_{A'} = 1 - f_A$. Containment, defines that A is a subset of B if $f_A \leqq f_B$. The union of two fuzzy sets is defined as the maximum membership of each set for each point and the intersection is defined as the minimum membership. Alternatively, the union/intersection of A and B can be defined as the smallest/largest fuzzy set containing/contained in both A and B.

Zadeh observed that the notion of "belonging" does not have the same role in fuzzy sets as in ordinary sets. It is only stated to be meaningful in the trivial sense of $f_A(x)$ being positive. Zadeh noted, however, that one could define levels $0 < \beta < \alpha < 1$ and declare that: (1) x belongs to A if $f_A(x) \geqq \alpha$ (2) x does not belong to A if $f_A(x) \leqq \beta$ and (3) x has an indeterminate status relative to A if $\beta < f_A(x) < \alpha$. "This leads to a three-valued logic with three truth values." Zadeh neglected to note that if we set $\alpha = \beta$ in the above, we can obtain two value logic with $\alpha = \beta$ as the threshold above which x belongs to A, and below it does not.

In his third chapter, Zadeh extended basic identities which hold for ordinary sets to fuzzy sets. De Morgan's laws and distributativity of union and intersection over each other are fairly intuitive. A way of interpreting unions and intersections was also proposed. Within the framework of ordinary sets, a set expressed in terms of a family of sets $A_1, \ldots, A_n$ and connectives $\cup$ and $\cap$, can be interpreted as a network of switches with $A_i \cap A_j$ and $A_i \cup A_j$ corresponding to series and parallel combinations. In fuzzy sets, sieves are used instead of switches. A quote is the most succinct way of describing it:

> "Specifically, let $f_i(x)$, $i = 1, \ldots, n$ denote the value of the membership function of A_i at x. Associate with $f_i(x)$ a sieve $S_i(x)$ whose meshes are of size $f_i(x)$. Then, $f_i(x) \vee f_j(x)$ and $f_i(x) \wedge f_j(x)$ correspond, respectively, to parallel and series combinations of $S_i(x)$ and $S_j(x)$."

Chapter Four was dedicated to algebraic operations. The algebraic product of A and B denoted by AB is defined in terms of the membership functions of A and B by $f_{AB} = f_A f_B$, and clearly $AB \subset A \cap B$. Unlike the algebraic product, the algebraic sum has an associated condition. The algebraic sum of A and B denoted by $A + B$ is defined by $f_{A+B} = f_A + f_B$, if and only if $f_A + f_B \leqq 1$ is satisfied for all x. The absolute difference of A and B denoted by $|A - B|$ is defined by $f_{|A-B|} = |f_A - f_B|$. The concept of a *relation* was also extended to *fuzzy relations*. While a relation in ordinary sets is defined as a set of ordered pairs, a fuzzy relation in X is a fuzzy set in the product space $X \times X$. Essentially each ordered pair has an associated degree of membership in the relation.

A notion called *convexity* was discussed in Chapter Five, but the material is outside the scope of this thesis, and so is omitted from this review. □

2.5 Tversky - Features of Similarity (1977)

In one publication examining similarity, Amos Tversky published Features of Similarity[34] in the *Psychological Review* Journal. In the paper, he examined "metric and dimensional assumptions underlying geometric representation of similarity." The major contribution of the paper however, is the development of a new process termed *feature-matching*. It is interesting to note that when the paper was written, similarity was

primarily cited as useful for individuals making classifications, forming concepts, and generalizing. Tversky wrote:

> *"the concept of similarity is ubiquitous in psychological theory. It underlies the accounts of stimulus and response generalization in learning, it is employed to explain errors in memory and pattern recognition, and it is central to the analysis of connotative meaning."*

At the present time, this statement contains renewed validity in the field of machine learning. An entire branch of research called *metric learning* has emerged which uses machine learning in an attempt to produce an optimized distance function between elements where the distance is the inverse of similarity [16]. Much of the research surrounding theoretical analysis of similarity relations had regarded dissimilarity as a metric distance between geometric points. Using this framework, Tversky references the following three axioms with regard to distance between points. (1) Minimality - the notion that an item has zero distance from itself, and some positive distance to all distinct items; (2) Symmetry - that one point is the same distance from a second point as the second point is from the first; and (3) the Triangle Inequality - which dictates that the distance directly from one point to another is always shorter or equal to the sum of the distances from each point to any intermediate point. While these axioms may be valid for geometric points, Tversky argued that they do not hold for similarity. For example, he noted that in recognition experiments objects are more frequently identified as other objects than as themselves, so if we take identification probability as a measure of similarity, then the axiom of minimality is violated, and thus shows the incompatibility with the distance model.

Tversky also takes the opportunity in this paper to point out a flaw in the traditional assumption that similarity is symmetric. He explained how similarity can be taken in the form "a is like b" which is directional. "...it has a subject a, and a referent, b, and it is not equivalent in general to the converse statement 'b is like a'."

Having provided explanations why the first two axioms are not valid in a psychological similarity setting, he focused next on the triangle inequality. Using similarity among countries as an example, he showed that even if one item is similar to a second for some reason, and that second item is similar to a third for another reason, the first and third items may not be at all similar. Not only does this express that a 'similarity' relation would not be transitive, but the triangle inequality is not valid either. It is easier to see by replacing similarity with distance. Contrary to intuition, items one and three may be farther apart than the sum of the distances from each item to the second item. Violation of the triangle inequality also shows that a geometric point representation of distance as an inverse to similarity is not always valid.

It seems he was using the idea that similarity can be measured in a variety of ways to justify his claims. He also cited variability of the reasoning for his justification. Since he interpreted similarity as a judgment by an individual, room must be left for subjectivity, and so absolute rules may not be applicable at all.

As an alternative, the approach proposed instead was termed *feature matching*. Given a universe of objects denoted by lower case letters, each is assumed to be represented by a set of attributes or features, the set of which is given by respective upper case letters. So A is the set of features of object a. Tversky then elaborated five assumptions he based his theory on. The first called *matching* expressed that similarity is a function on three inputs. The common features, and the two sets of

features unique to each object. The second, *monotonicity*, expresses that similarity increases with addition of common features and/or deletion of distinctive features. A function satisfying these two assumptions is called a matching function.

Regarding the third assumption, Tversky stated that *independence* was the major assumption associated with his theory. This showed the distance between each of two pairs of objects is the same if the difference between each pair of objects is the same set of features.

Two more assumptions were made, but rigorous formulations were left for the Appendix. *Solvability* required that the feature space be rich enough that certain similarity equations could be solved, and *invariance* ensured preservation of equivalence of intervals.

With these five assumptions, the representation theorem dictates that there exists some similarity scale S, and non-negative scale f, and defines the value of the former based on the value of the latter applied to the union and set differences of pairs of attribute sets. Called the *contrast model*, it expressed "similarity between objects as a weighted difference of the measures of their common and distinctive features, thereby allowing for a variety of similarity relations over the same domain."

Tversky next defined the *ratio model* which generalized set-theoretical models of similarity from the literature. The model essentially normalized the contrast model so that values lie between 0 and 1. It also specified parameters α and β to vary the value of elements unique to each set.

A review of the asymmetry in similarity models was a valuable portion of this publication. Tversky noted that though the model itself suggests that the similarity of A to B should be the same as B to A, there are indeed instances where they are deemed

to be different. The example given was pairs of countries. In an experiment Tversky conducted, an overwhelming majority of participants preferred to use the phrase, "North Korea is similar to Red China" over the alternative, "Red China is similar to North Korea" which demonstrated one example of asymmetry in similarity. A further asymmetry experiment also indicated that subjects preferred to judge similarities by comparing the more complex object as the referent to a simpler subject. In some sense, it seems that a complex object is more similar to a simple object than that simple object is to the same complex object.

Tversky also noted that Rosch [27] had also performed experiments supporting the notion that "prototypes" (taken as the more simple object) are preferred to be the item compared to. For example, subjects preferred "103 is virtually 100", to the phrase "100 is virtually 103". Tversky continued to support his asymmetry claim by discussing another study where Garner [8] asked subjects to select a pattern of dots similar but not identical to a given one. Subjects usually chose "good" patterns as responses to "bad" patterns, but the converse was rare. One thought from this author might be to dynamically define α and β according to which item being compared is more 'prototypical'. If these values are allowed to vary, even if asymmetry is allowed, it can be compensated for.

Tversky went on to give evidence supporting his hypothesis that similarity and difference are complements. This is of course a generalization, so Tversky noted that when assessing similarity subjects tend to assign more importance to the common elements, and the same is true for judging difference. Related to this notion, he also explained how similarity depends on context, and provided corresponding rationale. One strange consequence of Tversky's look at features in similarity, is that a pair of

objects with many common and many distinctive features can be judged to be both more similar and more different than another pair of objects with fewer similar and distinctive features. Experiments were also conducted supporting the claim that the universe of choices significantly impacts how similar objects are judged to be.

The paper concluded with a discussion of the role similarity plays in classification, and a brief look at comparisons in similes and metaphors. "It appears," Tversky wrote, "that people interpret similes by [...] scanning the feature space of the referent that are applicable to the subject." It is important to mention this limitation in the rough set setting. Without the symmetry axiom, there could be a case where $sim(A, B) > sim(B, A)$ which means that it is impossible to isolate an optimal approximation using any of the measures which are naturally symmetric, but the Tversky index allows for the exploration of properties of a non-symmetric similarity relation when $\alpha \neq \beta$. □

2.6 Tversky & Gati - Studies of Similarity (1978)

Amos Tversky and Itamar Gati wrote a chapter in *Cognition and Categorization* outlining their new theoretical analysis of similarity in which they also look at some empirical consequences[35]. They began their chapter by declaring how fundamental similarity is to life. Since, as they begin, "any event in the history of an organism is, in a sense, unique," when an organism demonstrates the ability to recognize, learn, and judge, it suggests that the organism can categorize stimuli and classify situations by similarity ([35],p.75).

Tversky and Gati observed that "the theoretical analysis of similarity relations [had] been dominated by geometric models," where each object is represented by a

point in some coordinate space, usually assumed to be Euclidean ([35],p.75). Analysis is generally done to embed the objects in a space of minimum dimensionality based on the observed similarities.

The authors then reviewed the feature-theoretical approach to analysis of similarity relations proposed by Tversky in [34]. This view was said to challenge the geometric approach as the measurement of similarity is quite different. It seems that the main difference was that the new feature-based approach did not try to restrict the number of dimensions at all. In some sense, although the authors do not discuss this explicitly, the new approach allows for any number of dimensions, if we regard each possible feature as a dimension and the possible attributes as the values in that dimension.

A brief comment on notation defined $s(a, b)$ as the observed similarity of a to b. It is expressed as a function of three arguments: $A \cap B$ (common features), $A - B$ (features of A but not B), and $B - A$ (features of B but not A). Also of note is that the feature-theoretical approach does not require subjective judgments based on functions of each feature like the euclidean approach, which must define the extent to which each feature should contribute to the final euclidean point.

The authors defined the *contrast model*, based on an interval similarity scale S, which preserves the observed similarity order, and a scale f defined on the relevant feature space such that $S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$ where $\theta, \alpha, \beta \geq 0$. They note that this defines a family of indicies defined by the three parameters rather than a unique index.

The remainder of the chapter analyzed problems using the contrast model and five studies they performed. The problems, all concern the impact on measured similarity

from the judgment of the *task* (similarity vs. difference), the *direction* of comparison (a vs. b or b vs. a), and the effective *context* (the whole set of objects). Actually, these are impacts that must be ignored in the main thesis contribution because of the rough set setting, and the need to begin with a simple theory that can be revised and improved.

Judgments of similarity and judgments of difference can be regarded as conceptually independent, but the authors cited that previous data appears to support the notion that they are perfectly correlated in many-though not all-cases. The authors explain that "the instruction to consider similarity may lead the subject to focus primarily on the features that contribute to the similarity of the stimuli, whereas the instruction to consider difference may lead the subject to focus primarily on the features that contribute to the difference between stimuli" ([35],p.77-78). They performed a study asking subjects to either judge the more similar or the more different pairs from two pairs of countries. The results obtained supported their claim (with $t = 3.27, df = 59, p < .01$).

The discussion on symmetry was continued from Tversky's previous paper [34], as the authors explain that "similarity judgments can be regarded as extensions of similarity statements" (e.g. "A is like B") ([35],p.80), and that this statement is directional. Thus more weight would be given to the features of the subject than the features of the referent which corresponds to $\alpha > \beta$ within the contrast model. It was shown algebraically that $s(a,b) > s(b,a) \iff f(B-A) > f(A-B)$ implying "that the direction of symmetry is determined by the relative salience of the stimuli so that the less salient stimulus is more similar to the salient stimulus than vice versa." ([35],p.81) The authors ran an experiment in which two groups participants were

asked to assess how different (or similar) countries are from one another, with each group getting the countries in opposite orders. The results supported the authors predictions.

While the first two problems were related to the parameters (θ, α, β), the third problem, dealing with contextual effects, describes how the function f is impacted by changing context. The example given, from when Germany was divided, was that East Germany and West Germany may be judged as very similar in a geographical or cultural context, but vastly different in a political context. Among a group of Asian and African countries the two Germanys would likely be viewed as more similar than among a group of European countries. An experiment was run to test the hypothesis that the same pair of objects are judged more similar among a larger group of varying objects with no commonalities, than among a smaller group of objects with features in common. Countries were used as objects, with their geographic continent as the common feature. Results the authors obtained, and results they cited from a paper written by Sjoberg [31], supported their conclusion. It seems that a complete measure of similarity should not ignore context though in certain settings the data are not provided to be able to consider it.

The context issue Tversky and Gati raised is one which should be explored in the rough set setting. The author observes that adding the same feature to all objects in a universe would skew similarity evaluations which are based on cardinality alone. Thus, it would seem that to provide a fully accurate measure of similarity, not only the objects being compared must be evaluated, but all objects in the universe under consideration. The extent to which objects not being compared to one another should be evaluated is an open problem (for example, they may just be checked to have

certain features or not, thus skipping some parts of a full evaluation). This adds several layers of complexity, so exactly how changing context could be incorporated into the measurement of similarity is an open problem for future work. It must be noted though, that by generalizing the distance function using measure theory, one type of context in our evaluation of similarity is inherently present, since the particular attributes of an object to measure and compare can be selected.

Finally, one of Tversky and Gati's conclusions should be reiterated. They reminded the reader that "there is no unitary concept of similarity that is applicable to all different experimental procedures used to elicit proximity data." With regard to the present work, this means that to increase the accuracy of any similarity measure, more functions and possibly more variables are needed to capture more information about the subject, referent, universe, and type of comparison being made. □

2.7 Tversky & Gati - Similarity, Separability, and the Triangle Inequality (1982)

In this *Psychological Review* article, Tversky and his coauthor Gati examined the triangle inequality with respect to similarity in several contexts[36]. The triangle inequality $D(i,j) + D(j,k) \geq D(i,k)$ algebraically dictates that total distance from one object to a third increases (or stays constant) if instead of finding the direct distance, we instead evaluate it through an intermediate second point. Inherently visually obvious in geometry, we can simply draw any triangle and observe that the lengths of two sides of a triangle cannot sum to less than the length of the third side. The degenerate case where two sides equal the third results in a line instead

of a triangle. Somewhat counter-intuitively this property is not guaranteed in all algebraic settings.

The authors of the paper restrict their attention to monotone proximity structures thus requiring three elementary ordinal properties: dominance, consistency, and transitivity. These are abstracted algebraic versions of properties we can visualize geometrically in a right triangle. Dominance simplifies in the context of a right triangle to simply mean that the longest side is longer than either of the other two, or as the authors write: "a two-dimensional difference exceeds its one-dimensional components." Consistency, meaning dimensions are assumed to be independent, can be visualized geometrically by noting that any line can be translated linearly, and the x and y components maintain the same ordering. Their third condition, transitivity, is required of a betweenness relation on pairs. It is thus the only example here which refers to four points of reference. To visualize a simplistic geometric example, we need to imagine four points. The property forces it to form a non-reflex quadrilateral (or two triangles sharing a single side if we connect either diagonal). If we label the farthest two points a and d, and the other two points b and c, we will find that the triangle formed by points a, b, c has longest edge (a, c), and the triangle formed by points b, c, d has longest edge (b, d). Given the above two conditions, transitivity states that the triangles a, b, d and a, c, d must have longest edge (a, d). Of course this is a simplification, but it serves to illustrate the conditions above for the Euclidean model. Tversky and Gati however, use the assumption that the distance between points $\delta(ap, bq)$ is related monotonically to the metric distance $D(ap, bq) = (|\dot{a} - \dot{b}|^{\gamma} + |\dot{p} - \dot{q}|^{\gamma})^{1-\gamma}$ where $\dot{a}, \dot{b}, \dot{p}, \dot{q}$ represent the one dimensional components or coordinates of each respective element, and $\gamma \geq 1$. This model is called

the *Minkowski γ-metric*, and classifies families of distances functions for each value of $\gamma \geq 1$. The Euclidean model has $\gamma = 2$, and for $\gamma = 1$ the distance between two points will always be exactly the same as it would be calculated through intermediate points. It is interesting to observe non-metric models where $\gamma < 1$. As this variable increases so does the distance saved in the model by computing directly between points instead of through an intermediate. For fractional values of γ in some abstract sense, the shortest distance between two points is not a straight line.

Metric models must also satisfy a property Tversky and Gati cited from (Beals et al., 1968) called segmental additivity, which breaks down to requiring that distance between any two points can be found. A condition called the *corner inequality* requires that the sum of the distances in each dimension is greater than or equal to the distance taken across all dimensions. They are equal only if $\gamma = 1$.

To contrast this model the authors outlined an approach based on feature matching and the contrast model previously proposed by Tversky[34]. A valuable mention of how the authors chose to represent quantitative attributes such as length or loudness, with smaller values given as subsets of larger values, was attributed to Guttman (1954), and Restle (1959,1961). The authors go on to define these nested sets in which smaller sets share all attributes of larger sets, and contrast them with chains of sets which share some but not all attributes. Of note, is their *coincidence hypothesis* in the contrast model which contradicts the corner inequality whenever more weight is given to common attributes of elements than to attributes unique to either set. The remaining contents of the article pertain to experiments illustrating the above properties. The authors asked subjects to rate (dis)similarity in many studies with different controlled variables, and analyzed the resulting data with their proposed equations.

The data from subjects who were asked to rate similarity between pictures of objects which had specific differences, strongly violated the corner hypothesis. This implies that when a sequence of pictures is arranged so that the difference between them is incremental and equal between adjacent pictures, subjects still perceive non-adjacent pictures as more similar than the sum of all similarities in adjacent pairs of picture in the chain. Only for colours, when hue and chroma are tested, is the corner equality strongly supported.

When the corner equality does not hold, the authors noted that while more dimensions could be introduced to compensate, in their experiments a third dimension did not fit data better than two dimensions, and so they suggest that dominance, consistency, and transitivity are not expected to hold in three dimensions. The corner equality is satisfied less often in certain conditions such as when attributes are more separable, when the structure of the dimensions is more obvious, or when common attributes of stimuli are weighted more than distinctive features. Stimuli are separable if one attribute can be examined ignoring the other. For example, color and shape are considered separable, while chroma and hue are not.

Tversky and Gati recall from earlier work that subjects tend to focus more on common attributes than distinct ones when judging similarity. This fact could be integrated into a similarity equation by requiring that attributes common to both elements be regarded as more important.

Strangely, when they studied proximate distance, they found that when items coincide on one dimension, bringing them closer increases their similarity, but when items differed in both dimensions, the same action decreases their perceived similarity.

Additionally, they note that attributes common to all elements of the domain

convey no information, so it may be desirable to eliminate them. This observation could be applied in situations where several items being compared pairwise have common attributes which could be ignored to modify pairwise comparison values.

The coincidence effect implies that either distances along straight lines are not additive, or the shortest path between points is not a straight line. Thus the existence of such an effect has not fully been explained. Several possible interpretations have been given though. The authors cited Krumhansl(1978), as suggesting that spatial density around a point may increase measured distance. In this view, the distance along two adjacent sides of a rectangle could be shorter than the diagonal if there exists a higher density of points near the values along the diagonal. Tversky and Gati ran another experiment however, in which the results detracted from this proposal.

The paper concluded with the claim that since, for example, the triangle inequality and segmental additivity are sometimes shown to be incorrect, basic properties of the geometric model cannot be universally applied as valid principles of psychological similarity. □

2.8 Pawlak - Rough Sets (1982)

The article *Rough Sets*[21] in the International Journal of Computer and Information Sciences, is routinely cited in any work involving rough sets. It is often cited as the foundational paper of rough set theory as it is the first published article, though in it Pawlak states that he first introduced Rough Sets in his 1981 report of the same name. In this paper, Zdzisław Pawlak introduced the now well known concepts of upper and lower approximations. He formalized rough sets as a pair of crisp sets giving the upper and lower boundaries. One result being that the lower must be a

subset of the upper, and if the rough set is not also crisp, the subset is proper.

In addition to defining the upper and lower approximations, Pawlak also defined 30 laws governing them. The laws dictate how negation, union, intersection, and the boundary region relate to one another. As part of rough approximations, Pawlak defined what it means for sets to be roughly equal, roughly bottom-equal, or roughly top-equal. Roughly equal refers to two sets having the same upper and lower approximations, while the other two define cases where only one approximation is equal.

Even without equality, two sets without any equal approximations, can still be compared, or related with Pawlak's rough inclusion. A set X is roughly bottom-included (top-included) in Y if the lower (upper) approximation of X is contained in the lower (upper) approximation of Y and roughly included if it is both bottom-included and top-included.

A small section of the paper was devoted to expressing the "quality" of an approximation. Pawlak defined the accuracy of an approximation by a ratio. He divided the number of elements in the lower approximation by the number of elements in the upper approximation to obtain a value between 0 and 1. For any crisp set, the upper approximation equals the lower approximation, so the accuracy is 1. If our target set contains some part of every equivalence class, but does not contain every element in any class, the accuracy is 0.

Obviously there are limitations with this definition. For any non-empty set where the lower approximation is the empty set, the accuracy is always zero. This leads to counter-intuitive situations where two sets are not equal, have the same approximation of zero, yet the accuracy of both is equal (to zero). For two different sets with the same approximation the accuracies of each are expected to be different. □

2.9 Pawlak - Rough Sets and Fuzzy Sets (1985)

Only a few years after introducing rough sets, Pawlak wrote this short communication comparing rough sets to fuzzy sets[22]. First, he reviewed rough set definitions, including the upper and lower approximations, and the boundary region corresponding to what this thesis refers to as a border set. He also noted that the approximation space uniquely determines a topological space using the elementary sets as a base. This led to 11 intuitive properties. Among them are the fact that the lower approximation is a subset of the target set, which is itself a subset of the upper approximation. Also, the fact that the lower and upper approximations of the entire universe of elements is the universe itself, and the empty set is also its own upper and lower approximation. The remaining properties involved the intersection and union of sets and upper/lower approximations as well as results of taking the upper/lower approximation twice or taking both approximations in sequence.

Pawlak briefly reviewed the definition of fuzzy sets from Zadeh [39] then defined a membership function that evaluates if an element is in the lower approximation, the boundary region, or outside the upper approximation altogether. It was shown that this membership function cannot be extended to union and intersection of sets, by deriving a contradiction using some of the properties defined earlier.

Finally, he noted that "the concept of rough set is wider than the concept of fuzzy set." Rough sets can be reduced to fuzzy sets in the following scenario: given two sets, the union of their lower approximations is not merely a subset of the lower approximation of the union but they are exactly equal, and the upper approximation of the intersection of two sets is, instead of a subset, equal to the union of upper approximations.

Essentially, the reduction can only occur when the union and intersection of two fuzzy sets are equal to the union and intersection if the sets were regarded as rough, and degree of membership is ignored. Since Pawlak first analyzed both types of sets, there have been several comparative studies to investigate similarities and differences [37]. They seem to conclude that whether rough set theory is regarded as a deviation or an extension of classical set theory depends on the view adopted; namely set-oriented, or operator-oriented. □

2.10 Pawlak - Rough Sets Approach to Knowledge-Based Decision Support (1997)

For the European Journal of Operational Research, Pawlak produced a paper in which he described possible applications of rough set theory[24]. He explained that the indiscernibility relation is the mathematical foundation of rough set theory. It expresses that the information we have defines objects, but due to a lack of knowledge, we are unable to distinguish between certain objects or groups of objects. The principle of rough set theory is that if we are unable to distinguish objects, but still require further analysis on the entire set of objects, we can define each object by the group it is in, and perform analysis on the groups instead of on the individual objects. As Pawlawk wrote, "any set of all indiscernible (similar) objects is called an elementary set, and forms the basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as crisp (precise) set - otherwise a set is rough (imprecise, vague)."([24],P.48) Another way to envision a set which is rough or imprecise, is as a group of items where some item is similar or indiscernible

to an item not in the group. This very fact is what makes the set rough. Since we are unable to describe the group using only a set of indiscernible items or classes-which may significantly reduce the size of the description-we must instead use the items themselves.

One disadvantage of vague or rough concepts Pawlak noted, was that while precise concepts can be characterized in terms of information about their elements, this is not possible for vague concepts. So, a rough set can be replaced/approximated by a pair of precise concepts called the upper and lower approximations. The lower approximation includes all objects that must belong to the concept, and the upper approximation contains all objects which might possibly belong to the concept. The difference between them Pawlak called the boundary region.

According to the Pawlak, at the time of writing, the rough set method branched into disciplinary fields including machine learning, knowledge discovery, statistics, and inductive inference, but the interpretation of results lies outside the theory. This seems logical since there is usually a wide variety of possible explanations for numerical or set-theoretical results.

For a more general understanding, Pawlak chose to define the framework referring to an *information table*, or *attribute-value table* first, instead of the mathematical relation definition, which followed. In such a table, each possible group of attributes/columns divides the objects into groups or 'classes' which have the same attribute values. To refer to a class containing an element x, we reference $I(B)(x)$ (meaning the indiscernibility relation based on the set B of attributes) or just $B(x)$. Using the set of attributes B to partition the elements, the groups formed are called *B–elementary sets*.

Pawlak defined the framework formally as follows:

> "Let U be a finite set of *objects* - called the *universe* - and let A be a finite set of *attributes*. With every attribute $a \in A$ set of its *values* V_a is associated. Each attribute a determines a function $f_A : U \to V_a$. With every subset B of attributes A we associate an indiscernibility relation on U, denoted $I(B)$ and defined thus: $I(B) = \{(x, y) \in U \times U : f_a(x) = f_a(y), \forall a \in B\}$" [24, p.49]

This definition allows for different sets of equivalence classes to be formed from the universe, by selecting different sets of attributes, and thus using a different indiscernibility relation. In this thesis, however, only one attribute or equivalence relation is considered so the universe is only divided up in one way, though it could be easily extended by restricting attention to a subset of attributes as Pawlak did here.

Pawlak went on to define the *B–lower* and *B–upper* approximations. These are simply versions of the lower and upper approximations which explicitly specify the group of attributes B being used to partition the universe. The *B–lower* approximation of a target set X includes every object class $B(x)$ where all objects in the class are within the set X. The *B–upper* approximation is similar but includes all object classes which contain one or more objects from the target set. Their difference is called the *B–Boundary region*.

Almost as a side note, Pawlak mentioned that a rough set can be characterized numerically by dividing the size (cardinality) of lower approximation by the size of the upper approximation to yield the coefficient $\alpha_B(X)$ which he called the *accuracy of approximation*. This coefficient could be used to check if a set X is crisp, since it is only crisp if $\alpha_B(X) = 1$. Values less than one obviously mean the set is rough.

Next the author defined rough membership. We recall the authors explanation that "vagueness is related to sets, while uncertainty is related to elements of sets" ([24],p.51) To discuss uncertainty in the rough set framework, Pawlak defined the *rough membership function* as $\mu_X^B(x) = \frac{|X \cap B(x)|}{|B(x)|}$. Seen as the degree of certainty to which x belongs to X, the function is simply a ratio which divides the number of objects common to X and the equivalence class containing x, by the total number of objects in the equivalence class containing x. A value of one shows that $X = B(x)$ so this function can also be interpreted as an evaluation of what portion of the objects indiscernible from x (the class $B(x)$) are in the target set X. Pawlak pointed out that values of the membership function are computed from given data, as opposed to being assumed like the fuzzy membership function.

The dependency of attributes was explored next, though the notion is not used in this thesis, and so is explored only briefly. If it is found that we can determine the value of a certain attribute for all objects in a universe using other attributes, we call this a *functional dependency* in the relational databases sense. In dealing with rough sets, Pawlak also defined *partial dependency* of attributes which evaluates the portion of the universe in which one group of attributes uniquely determines another group.

If we derive a total dependency, the dependent attributes could be considered unneeded since they can be uniquely determined by others. These attributes are called *superfluous* while all others are called *indispensable*. A set of attributes is independent (orthogonal) if all the attributes are indispensable. Pawlak defined a subset B' of B as a *reduct* of B if B' is independent and both partition the universe equally. The *core* of a set of attributes is the set of indispensable attributes. To connect the notions of

core and reduct, Pawlak reported that the core of a subset of attributes B is equal to the intersection of all reducts of B, i.e. $Core(B) = \bigcap Reducts(B)$.

Pawlak distinguished between two types of attributes: *condition* and *decision* attributes. If all attributes in an information table are classified as one of the above, the table becomes a *decision table*. Decision attributes specify decisions to be made based on condition attributes. When reducing condition attributes, one goal is to preserve the dependency between condition and decision attributes. Pawlak reported that it is easy to generalize the concept of a reduct so that features other than partitions, such as a degree of partial dependency, are preserved. He elaborated that if we regard condition attributes as a premise, and decision attributes as a conclusion, we can easily convert a decision table into a list of rules, with a 1–to–1 correspondence between rules and objects. Then, using a rough set technique, the rules can be merged to yield a minimal set of decision rules.

Decision problems involve a set of objects which could be actions, states, processes, opinions, or anything else. The goal of decision analysis is to explain a decision based on the cause of it being made, and to give guidelines how to make a decision depending on the situation. The rough set model provides techniques to create minimal sets of decision rules from decision tables, or a set of decision rules. To conclude, Pawlak cited numerous applications of the rough set theory, including medicine, engineering, and finance, and listed advantages of the rough set approach, including its simplicity and ability for data reduction. □

Chapter 3

Mathematical Background

3.1 Measures

In its most general sense, measure theory is quite literally the study of measurement and sizes. When defining an optimal set, we require a way to express multiple methods of measuring the size of a set. To think about this in a basic sense, imagine we might have a collection of a few large objects we wish to measure. If we measure the number of objects(cardinality), the size is small, but if we measure based on the total sizes, we get a different result. If we were comparing this set to a large group of small objects, the type measurement used will obviously impact how we judge these sets. The comparison could be made based on weight, colour, etc.

Measure theory essentially generalizes the concept of 'size'. When the size of an object or group of objects is measured, it can be done in various ways. For example, sometimes a measurement of length is enough (rarely do you hear about the width of a boat, because the length defines the size), but in other occasions width may be required so that the area can be calculated such as defining the amount of floor space

in a room. In this case, the length and width may be important to ensure the room is not excessively long and narrow. In other cases, only the final calculated area is of any value such as when the fire department assigns a maximum occupancy. In still other cases, we may desire the additional dimension of height if we are seeking volume. These all refer to physical size in one, two, and three spatial dimensions respectively.

While spatial measurements are valuable in some circumstances, there are often other measurements, observations, judgments, or facts which can be expressed only in non-spatial terms. An item or group of items will have a physical size, but also properties such as cost, weight, or quantity. We could even use subjective judgments such as how strongly someone 'prefers' things. Measure theory allows us to abstract the type of measurement so that we can make claims about all types of measurements. The only fundamental rule in measure theory says that if everything in set A is also in set B then set B is larger or equal to set A in whatever way we are measuring (and strictly larger, if B also contains at least one item not in A). For example, if a person 'likes' a set of items a certain amount, we assume they 'like' any set with all these items at least as much. Then, since we will assume a null-free universe, the addition of a new (non-null) item to a set requires that the measurement increase, i.e. if we add more stuff to the set, we assume the measurement of the set will be 'bigger' on whatever scale we use. There are obvious counterexamples, such as when negative weights are permitted, but this thesis restricts attention to the properties required. In what follows, the word *size* is taken to mean the measurement of an element, i.e. the evaluation of μ.

Since the field of study is so vast, only some basic results from *measure theory* are

outlined which have been adapted for the purpose of this thesis (c.f. [9, 20]).

Given a (not necessarily finite) set U, define a function $\mu : 2^U \to \mathbb{R}$, ($\mathbb{R}$ is the set of real numbers) that satisfies the following properties:

1. for all $X \subseteq U$, $0 \leq \mu(X) < \infty$,
2. $\mu(\emptyset) = 0$,
3. if $X_i \subseteq U$ for $i = 1, \ldots, \infty$ and $X_i \cap X_j = \emptyset$ if $i \neq j$, then

$$\mu(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} \mu(X_i).$$

Descriptively, the measurement of anything cannot be negative and must be finite, the empty set always has size zero, and the size of anything is equal to the sum of the sizes of its parts.

Any function which satisfies the above three criteria is called a *finite measure* over 2^U, and a triple $(U, 2^U, \mu)$ is a *measure space* (c.f. [9, 20]).

One can show that μ also satisfies:

- for all $X, Y \subseteq U$, if $X \subseteq Y$ then $\mu(X) \leq \mu(Y)$,
- for all $X_i \subseteq U$, where $i = 1, \ldots, \infty$ (and X_i are not necessarily disjoint), we have

$$\mu(\bigcup_{i=1}^{\infty} X_i) \leq \sum_{i=1}^{\infty} \mu(X_i).$$

As just noted, the empty set must have size zero. From the perspective of measure theory, a set is called negligible when its size is insignificant enough to be ignored

in certain settings. In a rigorous sense, "define a subset A of U to be negligible if for each positive ϵ there exists a finite or countable collection $I_1, I_2, \ldots$ of intervals satisfying $A \subset \bigcup_k I_k$ and $\Sigma_k |I_k| < \epsilon$." [1]

It should be addressed that the triple used here does contain the power set of U which is of course derivable from U, but this convention is used to conform to the same format as the measure space that Halmos used. As mentioned in the review of his paper (Section 2.2), instead of 2^U, he assumed a σ-ring S of subsets of the universe with $\bigcup S = X$ [9].

All sets whose size is zero are clearly negligible. If a continuous universe is used, any set with a single element would be regarded as negligible, and thus any finite set would also be negligible. In the finite rough set setting, however any non-zero measurements can not be ignored.

In the standard version of measure theory, the property of null set freeness is not defined and not discussed (c.f. [9, 20]), so it is technically possible to have other items measuring zero. This creates a bit of mathematical trouble later so this thesis creates a specification for measure spaces. If and only if the measurement of a set is zero, that set is the empty set. With this change, we can now assume that there will be no elements in any negligible set, and if a set is negligible, we know it is the empty set. Formally, define a set X such that $\mu(X) = 0$ as *μ-null set* and all μ-null sets as *negligible*. Observe that cardinality, is a null set free measure since the only set with size zero, is the null set.

- A measure space $(U, 2^U, \mu)$ is *null set free* if the empty set, $\emptyset$, is the only μ-null set, i.e. if $\mu(X) = 0 \iff X = \emptyset$.

If a set U is finite, the definition of a measure can be simplified.

- From (3) of the measure definition, we have that if $X = \{x_1, \ldots, x_m\}$, then $\mu(X) = \mu(\{x_1\}) + \ldots + \mu(\{x_m\})$.

This means that for finite sets a measure can be defined element-wise, as $\mu : U \to \mathbb{R}$ and then just extend it for sets in a standard way as, for every $X \subseteq U$,

$$\mu(X) = \sum_{x \in X} \mu(x).$$

Assume that if a set U is finite, a measure μ is element-wise defined. Discrete probability is an element-wise defined measure, with $\mu(U) = 1$.

- If U is finite than a measure space is null-free if for every $x \in U$, $\mu(x) > 0$.

Again, the most simplistic measure of a set, cardinality, is an example of a null-free element-wise defined measure given by $\mu(x) = 1$ for all $x \in U$.

3.1.1 Algebraic foundation

Next, a brief algebraic background is provided. From Halmos second chapter, a ring of sets (in the *measure* sense) is defined as a non empty class R of sets which is closed under unions and differences. i.e. if $E \in R$ and $F \in R$ then $E \cup F \in R$ and $E - F \in R$. Similarly, an algebra is a ring of sets which is also closed under union and complement operations[9]. Throughout this work, when we write that we have an algebra A, this is to what we are referring. Essentially, a collection of distinct entities, which may or may not overlap, but if they do, the overlap is clearly defined.

3.2 Rough Set Foundations

When Rough Sets were introduced by Pawlak, he credited fuzzy set theory proposed by Zadeh[39] as the most successful approach to the problem of imperfect knowledge[21]. While he admitted an overlap with many other theories, this thesis will follow Pawlak's initial work, and consider rough set theory as an independent discipline.

First some terms must be defined. To begin, imagine that we have group of items, alternatively called a space, or universe, and use U to represent it. The items in the universe may be called elements, objects, entities, etc., and will be represented by lower class characters, usually x, and are sometimes enumerated (although no ordering is implied). The most basic concept at the heart of rough set theory is an indiscernibility relation.

Imagine the universe of elements is duplicated so now there are two copies of our original universe. Then compare every element in the first version to every element in the duplicate. If every element in the universe is different, the result should get only one match for each element, since it will not be possible to tell the difference between the two duplicate elements. These two items could then be called indiscernible, that is, we are unable to tell the difference. This is a basic assumption of our universe of elements, that they are all distinct and discernible from one another. Sometimes however, elements can be grouped either by some commonality, or by some other means. If then, we try to compare groups, but only compare based on one measurement, then multiple groups may be indiscernible.

For example, if a group is discussing options to eat, the number of choices is essentially infinite but they can be grouped by differences, similarities, or arbitrarily,

perhaps by first letter, or type of food. Any grouping we form could be considered an equivalence relation. Choices in the same group, should be exactly the same according to the criteria we used to group them. So then, if a specific fast food chain is ruled out as an option because this chain is disliked by some, it is still discernible from other fast food chains. That is, we would not place it in the same group as fast food chains which are not disliked by anyone. But if a fast food chain is ruled out specifically because it is fast food, all other fast food chains are indiscernible in this situation. According to the situation at hand, all fast food chains are equivalent (in that they are excluded as choices). All similar choices, though not identical in every way, can still be said to be equivalent as they would be ruled out for the same reason. This group's opinions and choices could be said to partition the universe of elements into equivalence relations.

In general, if there is a finite non-empty universe of elements U, then let $E \subseteq U \times U$ be an equivalence relation, denote the equivalence class E containing x as $[x]_E$, and use U/E to represent the set of all equivalence classes of E. This is a kind of quotient space, but it cannot be called a quotient group since no inverse operations are assumed. It is not even a quotient structure because there is no assumption that there exists some operation defined to take two elements and return a single element. To be precise, the set itself cannot even be called an algebraic structure also due to lacking any (non-trivial) binary operations.

Since each equivalence relation creates sets of items which are indiscernible from one another, this work often refers to 1an entire class without wishing to reference a specific item within it. We may also reference classes without wishing to make or imply assumptions about what might be inside them. If all sets of indiscernible

elements are considered classes, we are left with *elementary sets*, *components*, or *atoms* which can be interpreted as basic observable, measurable, or definable sets. Denote this space $\mathfrak{Comp} = U/E$, and the elements within it by bold symbols, and write for example $\mathbf{x} \in \mathsf{B} \subseteq \mathfrak{Comp}$.

A pair $\mathcal{AS} = (U, E)$ of universe and equivalence relation, is referred to as a *Pawlak approximation space*. When working with rough sets one must minimally have the universe of elements and an equivalence class partitioning them. These are assumed to be given in subsequent sections.

3.2.1 Upper and Lower Approximations

If a non-empty set X of items cannot be represented accurately by equivalence classes, it would be because some item is in X, but a different item in the same equivalence class is not. We call a set like this *non-definable*, *non-exact*, or *vague*. This is the situation around which, this thesis is based. In traditional rough set theory, only two approximations of X can be taken: the lower approximation and the upper approximation, denoted $\underline{\mathbf{A}}(X)$ and $\overline{\mathbf{A}}(X)$, respectively.

Informally, the lower approximation of X consists of all of the equivalence classes (components) which have *all* of its elements in X, while the upper approximation of X consists of all the equivalence relations (components) that have *any* of its elements in X. An alternative way of looking at this is if an equivalence class from the lower approximation is provided, it is guaranteed that every element of that class is a part of the lower approximation. Alternatively, if an equivalence class from the upper approximation is provided, we can only guarantee that at least one element from the class is part of the upper approximation. A formal definition follows:

Definition 1 ([21, 23])**.** *For each $X \subseteq U$,*

1. $\underline{\mathbf{A}}(X) = \bigcup\{\mathbf{x} \mid \mathbf{x} \in \mathfrak{Comp} \wedge \mathbf{x} \subseteq X\}$,
2. $\overline{\mathbf{A}}(X) = \bigcup\{\mathbf{x} \mid \mathbf{x} \in \mathfrak{Comp} \wedge \mathbf{x} \cap X \neq \emptyset\}$. □

It is obvious that all elements in the lower approximation of X, are in the set X, and all elements in the set X are in the upper approximation of X. i.e. $\underline{\mathbf{A}}(X) \subseteq X \subseteq \overline{\mathbf{A}}(X)$. It is also important to mention the existence of many extensions and versions of this basic model [12, 38, 32, 25, 6].

A set $A \subseteq U$ is *definable* (or *exact*) [21] if it is a union of some equivalence classes of the equivalence relation E. Let $\mathbb{D}$ denote the family of all definable sets defined by the space (U, E). Formally

$$\mathsf{A} \in \mathbb{D} \iff \exists \mathsf{C} \subseteq \mathfrak{Comp}.\ \mathsf{A} = \bigcup_{\mathbf{x} \in \mathsf{C}} \mathbf{x},$$

or, equivalently, as the universe U is finite,

$$\mathsf{A} \in \mathbb{D} \iff \exists \mathbf{x}_1, \ldots, \mathbf{x}_n \subseteq \mathfrak{Comp}.\ \mathsf{A} = \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_n.$$

We would like to point out the duality of $\mathfrak{Comp}$ and $\mathbb{D}$. Each set of components $\mathsf{C} \subseteq \mathfrak{Comp}$ uniquely defines the *definable set* $\mathrm{dset}(\mathsf{C}) \in \mathbb{D}$, as $\mathrm{dset}(\mathsf{C}) = \bigcup_{\mathbf{x} \in \mathsf{C}} \mathbf{x}$, and each definable set $\mathsf{A} \in \mathbb{D}$ uniquely defines the *set of components* $\mathrm{comp}(\mathsf{A}) \subseteq \mathfrak{Comp}$, by $\mathrm{comp}(\mathsf{A}) = \{\mathbf{x} \mid \mathbf{x} \subseteq \mathsf{A}\}$.

Moreover, for each set of components $\mathsf{C} \subseteq \mathfrak{Comp}$, $\mathrm{comp}(\mathrm{dset}(\mathsf{C})) = \mathsf{C}$, and for each definable set $\mathsf{A} \in \mathbb{D}$, $\mathrm{dset}(\mathrm{comp}(\mathsf{A})) = \mathsf{A}$.

It follows that every lower and upper approximation is a definable set, i.e. $\underline{\mathbf{A}}(X) \in \mathbb{D}$ and $\overline{\mathbf{A}}(X) \in \mathbb{D}$ for every $X \subseteq U$. Furthermore, all definable sets are equal to their lower and upper approximations, as the corollary below shows.

Corollary 1. *For every* $X \subseteq U$, $X \in \mathbb{D} \iff \underline{\mathbf{A}}(X) = \overline{\mathbf{A}}(X) = X$. □

3.2.2 Borders

Since the definable sets in the area between the upper and lower approximations will play an important role in our model, we need to precisely define this area.

Definition 2. *For every* $X \subseteq U$, *we define the set of components* $\mathfrak{B}(X) \subseteq \mathfrak{Comp}$ *called the* ***border*** *of* X, *and the set of* ***border sets*** *of* X *called* $\mathbb{B}(X) \subseteq \mathbb{D}$, *as follows:*

1. $\mathbf{x} \in \mathfrak{B}(X) \iff \mathbf{x} \in \text{comp}(\overline{\mathbf{A}}(X)) \setminus \text{comp}(\underline{\mathbf{A}}(X))$,
2. $\mathsf{A} \in \mathbb{B}(X) \iff \mathsf{A} \subseteq \overline{\mathbf{A}}(X) \setminus \underline{\mathbf{A}}(X) \wedge \mathsf{A} \in \mathbb{D}$. □

The border and border sets (or boundary as Pawlak called it[21]) are building blocks for the optimal approximation defined later. The corollary below describes basic properties of borders and border sets.

Corollary 2. *For every* $X \subseteq U$,

1. $\text{dset}(\mathfrak{B}(X)) = \overline{\mathbf{A}}(X) \setminus \underline{\mathbf{A}}(X) \in \mathbb{B}(X)$ *and* $\mathfrak{B}(X) \subseteq \mathbb{B}(X)$,
2. $\mathsf{A} \in \mathbb{B}(X) \iff \exists \mathbf{x}_1, \ldots, \mathbf{x}_n \subseteq \mathfrak{B}(X).\ \mathsf{A} = \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_n$,
3. *if* $\mathsf{A} \in \mathbb{B}(X)$ *then* $\mathsf{A} \cap X \neq \emptyset$ *and* $\mathsf{A} \setminus X \neq \emptyset$.
4. *if* $X \in \mathbb{D}$ *then* $\mathbb{B}(X) = \emptyset$. □

Corollary 2(3) will often be used later in proofs of many important results of this thesis. It states that any element (set) in the set of border sets overlaps, or shares at least one element with the target set, and itself contains an item which is not part of the target set. This makes sense, otherwise the set would be outside the upper approximation or entirely contained in the lower approximation, which means in either case it would not be part of the set of border sets. Combined with corollary 2(4) they state that if X is not definable, then it overlaps with each element of its set of border sets.

Chapter 4

Similarity

From the chapter on measures and Measure Theory, recall that there are many ways to judge size. There are, likewise, various alternatives to measuring how alike two sets of items are. The amount of similarity can be regarded as a form of size, and so it has various methods of measurement that can be used. Among these possibilities, there must be some constant properties and this thesis refers to them as *axioms*.

4.1 Axioms

To express the axioms, assume that we have a set U (not necessarily finite) and a finite measure space $(U, 2^U, \mu)$. Suppose that we have a (total) function $sim : 2^U \times 2^U \to [0, 1]$ that measures *similarity* between sets. These types of functions have been known since at least the beginning of the twentieth century [11], but they do not have standard indisputable axiomatization [4]. Depending on the area of application, some desirable properties may vary [4, 26, 34].

This work assumes that any valid similarity function sim satisfies the following

five, intuitive axioms. Namely, for all sets $A, B \subseteq U$, we have:

S1 (Maximum): $sim(A, B) = 1 \iff A = B$,

S2 (Symmetry): $sim(A, B) = sim(B, A)$,

S3 (Minimum): $sim(A, B) = 0 \iff A \cap B = \emptyset$,

S4 (Inclusion): if $a \in B \setminus A$ then $sim(A, B) < sim(A \cup \{a\}, B)$,

S5 (Exclusion): if $a \notin A \cup B$ and $A \cap B \neq \emptyset$ then $sim(A, B) > sim(A \cup \{a\}, B)$.

We have also proposed a weakened version of S5, namely:

S5′ (Weak Exclusion) : if $a \notin B$ then $sim(A, B) \geq sim(A \cup \{a\}, B)$

The first axiom, *maximum similarity*, simply forces a similarity measurements to be at most 1. This can be easily achieved for any measure by rescaling values as a fraction of the highest possible value. Consider for example, a measurement function that returned values between 0 and 50, it could be scaled to divide all returned values by 50 so that the maximum similarity axiom is satisfied.

The *symmetrical similarity* axiom is a property we assume because of our context, but it is admittedly not valid in all practical cases, and the Tversky index was, in a way, created in objection to the axiom being assumed[34]. It is, however, required to obtain a metrical measure. In one sense, because the rough set setting is a context where dealing with having limited information is essentially fundamental, it is natural to assume that the information given does not make it possible to differentiate

between how similar A is to B, and how similar B is to A. Another aspect that must be mentioned is the assumption that *objective* similarity is being measured, not subjective judgments of how similar one set is to another, as in Tversky. This thesis must then assume the 'dimensionally organized space' that Tversky objected to in [34].

The third axiom, called *minimum similarity*, when coupled with S1, organizes the possible similarity values into the range $[0, 1]$. It is meant to define when sets are least similar, and it ensures that two sets are not minimally similar when they have any items in common. Without S3, there might be a case with some measure where two sets have a common element and have similarity zero. In this case, the first set can be duplicated only removing the common element(s) and this new set should be less similar to the second than the original. It should be less similar because it now has fewer common elements, but to be less than zero, the value would need to be negative. To correct this in general, the measure is scaled by subtracting the smallest possible returned value from every returned value. Most indexes assume the axioms S1–S3 either explicitly or implicitly, as any specific index could easily be scaled to fit S1 and S3, and S2 is required to obtain a metric space from the approximation values where distance between two points is a valid measurement of how dissimilar they are.

The axioms S4 and S5, although satisfied by many known similarities, were only explicitly proposed recently in [13]. They deal with changing sizes of sets. Here they are called *monotonicity axioms*. The axiom S4, referred to as *inclusion* is intuitive. If part of B is added to A, the result is closer to B than A alone. This is expected since A now contains more of B than it did before, so it should be more similar to it. The axiom S5 however, produces a few more issues. It can be reduced to the notion

that if some *new* element not in B is added to A, then the result is more different from B than A alone. This makes sense, since adding something that is not in either set clearly should make them less similar.

Notice the restriction to axiom S5, so it is only applicable when the sets being compared have at least one common element. If these two sets were mutually exclusive, there may be a case where $sim(A, B) = sim(A \cup \{a\}, B) = 0$ which violates the 'greater than' inequality. This leads to a weakened fifth axiom, in which adding an element to A which was not in B, may decrease the similarity between them, but not necessarily. Since the weakened version, S5′ allows for equality, the common element requirement may also be removed. We still must require the element not to be a member of set B, or adding it to A would make it more similar. Note, that if $a \in A$ the axiom is trivially true. Thus S5′ requires only $a \notin B$ instead of $a \notin A \cup B$. All the axioms formulated above follow from [13].

Recall that a measure of similarity *sim* is said to be *metrical* (i.e. it is a suitable tool to evaluate distance between two sets), if the function $diff(A, B) = 1 - sim(A, B)$ is a proper *metric* or *distance* which holds for all $A, B \subseteq U$. For a distance function f to be metrical, it must satisfy the following four requirements: [3, 4]

1. $f(A, B) \geq 0$,
2. $f(A, B) = 0 \iff A = B$,
3. $f(A, B) = f(B, A)$,
4. $f(A, C) \leq f(A, B) + f(B, C)$, i.e triangle inequality.

The first axiom can actually be neglected, since it follows from the other three. These axioms are those which govern metric spaces. According to [29], Metric Spaces were

originally proposed in 1906 by M. Fréchet in [7].

4.2 Similarity Indexes

The first similarity measure was proposed in 1901 by P. Jaccard [11]. It is still one of the most popular, possibly due to its simplicity. Notice that it only uses the magnitudes of union and intersection, and they are each only used once with no mathematical modifications. There are, however, other similarity measures which are still prominent in the literature. They are used, and were developed, for varying reasons which were investigated in Chapter 2.

- *Jaccard index* [11]: $sim_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$,
- *Dice-Sørensen index* [5, 33]: $sim_{DS}(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} = \frac{2|X \cap Y|}{|X \cup Y|+|X \cap Y|}$,
- *Marczewski-Steinhaus μ-index* [18, 19]: $sim_{MS}(X,Y) = \frac{\mu(X \cap Y)}{\mu(X \cup Y)}$, where μ is a *finite measure* on some U such that $X, Y \subseteq U$,
- *Tversky index* [34]: $sim_T^{\alpha,\beta}(X,Y) = \frac{|X \cap Y|}{|X \cap Y|+\alpha|X \setminus Y|+\beta|Y \setminus X|}$,
  where $\alpha, \beta \geq 0$ are parameters. Note that for $\alpha = \beta = 1$, $sim_T^{\alpha,\beta}(X,Y) = sim_J(X,Y)$ and for $\alpha = \beta = 0.5$, $sim_T^{\alpha,\beta}(X,Y) = sim_{DS}(X,Y)$.
- *Braun-Blanquet index*[2, 4]: $sim_{BB}(X,Y) = \frac{|X \cap Y|}{\max(|X|,|Y|)}$.

The Jaccard index is a simple fraction of common elements divided by the total number of elements. The quite similar Dice-Sørensen (DS) index adds a slight bit of complexity due to the additions involved, but uses essentially the same tools. It divides twice the number of common elements by the sum of the number of elements in

each set. If the sizes of each set (for whatever reason) are more difficult to determine or work with algebraically than the sizes of the intersection and union, notice that an equivalent denominator is the sum of the sizes of the union and intersection of the sets. In the DS index, those elements in both sets are counted twice in both the numerator and denominator. Simple algebra on the measures can show that the DS index will produce larger values than the Jaccard index whenever the union of the sets is larger than intersection, i.e. in all cases except trivially when the sets are equal, so both measures return value one.

The Marczewski-Steinhaus (MS) μ-index is a more generalized version of the Jaccard index. If $\mu(X) = |X|$, then $sim_J(X,Y) = sim_{MS}(X,Y)$ showing the Jaccard index is a special case of the Marczewski-Steinhaus μ-index. If instead of using the number of elements in the union and intersection, the total combined mass, volume, surface area, or cost of all elements in the union and intersection was used, it would require using the Marczewski-Steinhaus index rather than the Jaccard index. Given that there are a multitude of ways to measure size, it is quite helpful to note this generalization. We will develop our algorithm using the Jaccard index, but it is valuable to be able to generalize all formulas to work with the MS index so other types of size, rather than simple counting of elements, can be used.

The Tversky index, as detailed in Section 2.5, is similar to the Jaccard index, but makes it possible to define the weight of elements exclusive to each set separately. The Braun-Blanquet (BB) index was recently reinvented and analyzed in [26] in the context of Fuzzy Sets. [1] The index divides the number of common elements by the size of the larger set. This index does indeed maintain maximum and minimum axiomatic properties, although the values returned are larger than Jaccard or DS

[1] The authors were probably unaware of its long existence.

indexes, since by definition the denominator is smaller. Compared to the Jaccard index, the denominator of the BB index excludes the elements unique to the smaller set.

It is also useful to mention a special version of the Tversky index, when $\alpha = \beta$. Refer to it as the *Symmetric Tversky index*, and define it formally as

- *Symmetric Tversky index*: $sim^{\alpha}_{sT}(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X \setminus Y| + \alpha|Y \setminus X|} = \frac{|X \cap Y|}{|X \cap Y| + \alpha|(X \cup Y) \setminus (X \cap Y)|}$.

Similarly as the Jaccard index can be generalized by substituting *finite measures* instead of cardinality, the same can be done with the other indexes to create the following corresponding μ-indexes:

- *Dice-Sørensen μ-index* : $sim_{\mu DS}(X,Y) = \frac{2\mu(X \cap Y)}{\mu(X) + \mu(Y)}$,
- *Tversky μ-index*: $sim^{\alpha,\beta}_{\mu T}(X,Y) = \frac{\mu(X \cap Y)}{\mu(X \cap Y) + \alpha\mu(X \setminus Y) + \beta\mu(Y \setminus X)}$,
- *Symmetric Tversky μ-index*:
$$sim^{\alpha}_{\mu T}(X,Y) = \frac{\mu(X \cap Y)}{\mu(X \cap Y) + \alpha\mu(X \setminus Y) + \alpha\mu(Y \setminus X)} = \frac{\mu(X \cap Y)}{\mu(X \cap Y) + \alpha\mu((X \cup Y) \setminus (X \cap Y))},$$
- *Braun-Blanquet μ-index*: $sim_{\mu BB}(X,Y) = \frac{\mu(X \cap Y)}{\max(\mu(X), \mu(Y))}$.

So, based on the terminology introduced above, the Jaccard μ-index and the Marczewski-Steinhaus μ-index are the same formula.

All the similarity indexes above clearly have values between 0 and 1 and all, except (general) Tversky index and Tversky μ-index, satisfy the similarity axioms $S1$–$S4$. The Tversky index is not symmetric, in general, so it may not satisfy $S2$. The Tversky index is an asymmetric (by design) similarity index on sets that compares a variant to a prototype. If we consider X to be the prototype and Y to be the variant, then α corresponds to the weight of the prototype and β corresponds to the weight of the

variant. For the interpretation of X and Y as prototype and variant, α usually differs from β [34]. However for the interpretations used in this thesis, the case $\alpha \neq \beta$ does not make much sense. The Jaccard and Dice-Sørensen indexes satisfy S5, but the Braun-Blanquet index only satisfies the weaker axiom S5′.

Proposition 1 (Similarity Axioms and Similarity Indexes)**.**

1. *Marczewski-Steinhaus μ-index, Dice-Sørensen μ-index and Symmetric Tversky μ-index satisfy axioms $S1$–$S5$.*

2. *Tversky μ-index satisfies axioms $S1$ and $S3$–$S5$. It satisfies $S2$ if and only if $\alpha = \beta$. Tversky index also satisfies $S2$ if and only if $\alpha = \beta$.*

3. *Braun-Blanquet μ-index satisfies axioms $S1$–$S4$ and $S5'$.*

Proof. In the proof below we require a finite Universe, and a measure μ which must be finite and null-free. Note that none of our results are guaranteed to hold if the measure μ is not finite and null-free.
(*Marczewski-Steinhaus μ-index*) For S1, $sim_{MS}(A, B) = 1 \iff \frac{\mu(A \cap B)}{\mu(A \cup B)} = 1 \iff \mu(A \cap B) = \mu(A \cup B) \iff A = B$.

The reflexivity axiom $S2$ is trivial since both intersection and union of sets are commutative. Put algebraically: $sim_{MS}(A, B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} = \frac{\mu(B \cap A)}{\mu(B \cup A)} = sim_{MS}(B, A)$.

For S3, $sim_{MS}(A, B) = 0 \iff \frac{\mu(A \cap B)}{\mu(A \cup B)} = 0 \iff \mu(A \cap B) = 0$. Since any set with a measure of zero must be the empty set (due to null set freeness), we have that $A \cap B = \emptyset$.

The axioms $S4$ and $S5$ require a bit of algebra. If $a \in B \setminus A$, then $a \notin A \cap B$, so $\mu((A \cup \{a\}) \cap B) = \mu(A \cap B) + \mu(\{a\})$. On the other hand $A \cup B = (A \cup \{a\}) \cup B$, so $sim_{MS}(A, B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} < \frac{\mu(A \cap B) + \mu(\{a\})}{\mu(A \cup B)} = \frac{\mu((A \cup \{a\}) \cap B)}{\mu((A \cup \{a\}) \cup B)} = sim_{MS}(A \cup \{a\}, B)$. Hence

$S4$ does hold.

If $a \notin A \cup B$ then $A \cap B = (A \cup \{a\}) \cap B$ and $(A \cup \{a\}) \cup B = (A \cup B) \cup \{a\}$, so $\mu((A \cup \{a\}) \cup B) = \mu(A \cup B) + \mu(\{a\})$.

Thus, $sim_{MS}(A, B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} > \frac{\mu(A \cap B)}{\mu(A \cup B) + \mu(\{a\})} = sim_{MS}(A \cup \{a\}, B)$, so $S5$ holds too.

(*Dice-Sørensen μ-index*) For S1, $sim_{DS}(A, B) = 1 \iff \frac{2\mu(A \cap B)}{\mu(A) + \mu(B)} = 1 \iff 2\mu(A \cap B) = \mu(A) + \mu(B) \iff A = B$. This is due to the set-theoretical fact that $X = Y \iff X = Y = X \cap Y$ and all items must have positive measure weight.

Axiom S2 is again trivial because both set-intersection and addition are commutative. So $sim_{DS}(A, B) = \frac{2\mu(A \cap B)}{\mu(A) + \mu(B)} = \frac{2\mu(B \cap A)}{\mu(B) + \mu(A)} = sim_{DS}(B, A)$.

For S3, $sim_{DS}(A, B) = 0 \iff \frac{2\mu(A \cap B)}{\mu(A) + \mu(B)} = 0 \iff 2\mu(A \cap B) = 0 \iff \mu(A \cap B) = 0$ and since we are assuming that μ is a null-free measure, $A \cap B$ must be empty.

For $S4$, as in the MS index, we have again $\mu((A \cup \{a\}) \cap B) = \mu(A \cap B) + \mu(\{a\})$. Define $n = \mu(A \cap B)$, $m = \mu(A) + \mu(B)$. Since we have an element in B which is not in A, we know that $A \neq B$ so clearly $n < m$. Also notice that due to a being disjoint with A, $\mu(A \cup \{a\}) = \mu(A) + \mu(\{a\})$. Hence: $n < m \iff 2nm + 2n\mu(\{a\}) < 2nm + 2m\mu(\{a\}) \iff sim_{\mu DS}(A, B) = \frac{2n}{m} < \frac{2n + 2\mu(\{a\})}{m + \mu(\{a\})} = \frac{2\mu((A \cup \{a\}) \cap B)}{\mu(A \cup \{a\}) + \mu(B)} = sim_{\mu DS}(A \cup \{a\}, B)$, which means that $S4$ holds.

If $a \notin A \cup B$, and assuming $a \neq \emptyset$, then $sim_{\mu DS}(A, B) = \frac{2n}{m} > \frac{2n}{m + \mu(\{a\})} = sim_{\mu DS}(A \cup \{a\}, B)$, so $S5$ holds too.

(*Tversky and Symmetric Tversky μ-indexes*)[2] We begin for S1 with $sim_{\mu T}^{\alpha,\beta}(A,B) = 1 \iff \frac{\mu(A\cap B)}{\mu(A\cap B)+\alpha\mu(A\setminus B)+\beta\mu(B\setminus A)} = 1 \iff \mu(A\cap B) = \mu(A\cap B)+\alpha\mu(A\setminus B)+\beta\mu(B\setminus A) \iff \alpha\mu(A\setminus B) = \beta\mu(B\setminus A) = 0 \iff A = B$. Since weight cannot be negative, both set differences must have size measured to be zero, and because the measure is null-free, these set differences can only evaluate to the empty set. Only if two sets are equal are both set differences empty, and thus their similarity must have the maximum value of one.

The general Tversky μ-index does not satisfy S2 (due to possible assignments of α and β), but if we restrict attention to the symmetric version, it is trivial, due to commutative properties as above. Observe that $sim_{\mu T}^{\alpha,\alpha}(A,B) = \frac{\mu(A\cap B)}{\mu(A\cap B)+\alpha\mu(A\setminus B)+\alpha\mu(B\setminus A)} = \frac{\mu(B\cap A)}{\mu(B\cap A)+\alpha\mu(B\setminus A)+\alpha\mu(A\setminus B)} = sim_{\mu T}^{\alpha,\alpha}(B,A)$.

For S3, we return to using the general Tversky μ-index. Observe that $sim_{\mu T}^{\alpha,\beta}(A,B) = 0 \iff \frac{\mu(A\cap B)}{\mu(A\cap B)+\alpha\mu(A\setminus B)+\beta\mu(B\setminus A)} = 0 \iff \mu(A\cap B) = 0 \iff A\cap B = \emptyset$, again due to positive weights, and null set freeness.

To show S4, if $X = Y$ then $sim_{\mu T}^{\alpha,\beta}(X,Y) = sim_{\mu T}^{\alpha,\beta}(Y,X)$. If $X \neq Y$ then $sim_{\mu T}^{\alpha,\beta}(X,Y) = sim_{\mu T}^{\alpha,\beta}(Y,X) \iff \alpha = \beta$. Let $n = \mu(A\cap B)$, $k = \mu(A\setminus B)$ and $l = \mu(B\setminus A)$. If $a \in B\setminus A$ then $\mu(A\cup\{a\})\cap B = \mu(A\cap B)+\mu(\{a\}) = n+\mu(\{a\})$, $(A\cup\{a\})\setminus B = A\setminus B$, and $B\setminus A = (B\setminus(A\cup\{a\}))\cup\{a\}$. Hence

$$sim_{\mu T}^{\alpha,\beta}(A,B) = \frac{n}{n+\alpha k+\beta l} < \frac{n+\mu(\{a\})}{n+\alpha k+\beta(l-\mu(\{a\}))} = sim_{\mu T}^{\alpha,\beta}(A\cup\{a\},B),$$

so $S4$ must be valid.

If $a \notin A\cup B$, assign n, k, and l as above, then $sim_{\mu T}^{\alpha,\beta}(A,B) = \frac{n}{n+\alpha k+\beta l} > \frac{n}{n+\alpha(k+\mu(\{a\}))+\beta l} = sim_{\mu T}^{\alpha,\beta}(A\cup\{a\},B)$, so $S5$ holds too.

[2]Note that if the general Tversky μ-index satisfies an axiom, the symmetric index must also satisfy it (but the reverse is not necessarily true).

(*Braun-Blanquet index*) It is easy to show how it satisfies S1–S3. For S1, $sim(A,B) = 1 \iff \frac{\mu(A\cap B)}{\max(\mu(A),\mu(B))} = 1 \iff \mu(A\cap B) = \max(\mu(A),\mu(B))$ which can only be true if $A = B$. The second axiom can be shown using the same commutative reasoning as the previous indexes like so: $sim(A,B) = \frac{\mu(A\cap B)}{\max(\mu(A),\mu(B))} = \frac{\mu(B\cap A)}{\max(\mu(B),\mu(A))} = sim(B,A)$.

For S3, observe the following equivalence: $sim(A,B) = 1 \iff \frac{\mu(A\cap B)}{\max(\mu(A),\mu(B))} = 0 \iff \mu(A\cap B) = 0$. Due to null set freeness, the intersection must be the empty set, so this is proved.

The last two axioms are once again more involved. First, break them down into cases due to the max function. Validity must be proven for three cases: (1) $\mu(A) \geq \mu(B)$, (2) $\mu(A) + \mu(\{a\}) < \mu(B)$, and (3) $\mu(A) < \mu(B) < \mu(A) + \mu(\{a\})$. First, to show (1) let $n = \mu(A\cap B)$, $r = \mu(A)$ and $s = \mu(B)$. If $a \in B \setminus A$ and $\mu(A) \geq \mu(B)$, then $n < r$ (since the sets are not equal) so $sim_{\mu BB}(A,B) = \frac{n}{r} < \frac{n+\mu(\{a\})}{r+\mu(\{a\})} = sim_{\mu BB}(A\cup\{a\},B)$ and thus the first case holds. For (2) if $\mu(A)+\mu(\{a\}) < \mu(B)$, then $sim_{\mu BB}(A,B) = \frac{n}{s} < \frac{n+\mu(\{a\})}{s} = sim_{\mu BB}(A\cup\{a\},B)$, showing the second case is valid. Then for the final case (3), if $\mu(A) < \mu(B) < \mu(A) + \mu(\{a\})$, then $sim_{\mu BB}(A,B) = \frac{n}{s} < \frac{n+\mu(\{a\})}{r+\mu(\{a\})} = sim_{\mu BB}(A\cup\{a\},B)$ since $r < s$ so the third case is valid too. Hence $S4$ does hold in this case.

The last axiom for this index causes some trouble. Assign n, r, and s as above. For the first case, if $a \notin A\cup B$ and $\mu(B) \leq \mu(A)$, then $sim_{\mu BB}(A,B) = \frac{n}{r} > \frac{n}{r+\mu(\{a\})} = sim_{\mu BB}(A\cup\{a\},B)$. The second case is if $\mu(A) + \mu(\{a\}) > \mu(B) > \mu(A)$, then $sim_{\mu BB}(A,B) = \frac{n}{r} > \frac{n}{s+\mu(\{a\})} = sim_{\mu BB}(A\cup\{a\},B)$ because $s > r$. However, for the third case, if $\mu(B) > \mu(A) + \mu(\{a\})$, then $sim_{\mu BB}(A,B) = \frac{n}{s} = sim_{\mu BB}(A\cup\{a\},B)$. Though two cases hold, since they are equal in one case, this means that for the BB-index only S5′ is satisfied, and S5 is not. □

From all the indexes analyzed above, only the Marczewski-Steinhaus μ-index (i.e. also Jaccard index) is metrical since only the MS index (subtracted from 1) satisfies all metric space axioms[18]. Assuming we have a set X of items which can be regarded as 'points', for an index to be metric, it must satisfy the four conditions of a metric space[17, 9]:

1. For any two points, a positive real number is assigned (to represent distance)
2. The difference (distance) between two points is zero if and only if the points are equal
3. The difference from one point to a second point is the same as the difference from the second point back to the first
4. The triangle inequality is satisfied for all points: meaning that the distance directly between two points must be smaller than the sum of the distances from each point to a third.

Also $\mathit{diff}_{MS}(X,Y) = \frac{\mu((X\setminus Y)\cup(Y\setminus X))}{\mu(X\cup Y)}$, appears to have a natural interpretation, while the other differences, $\mathit{diff}_{\mu DS}(X,Y)$, $\mathit{diff}_{\mu T}^{\alpha,\beta}(X,Y)$, and $\mathit{diff}_{\mu BB}(X,Y)$ look rather artificial.

The Symmetric Tversky index and μ-index are useful when one wants to express the difference of importance (w.r.t. similarity) between the intersection $X \cap Y$ and the rest of $X \cup Y$, i.e. $(X \cup Y) \setminus (X \cap Y)$. If $\alpha < 1$, the measured size of $X \cap Y$ is more influential than that of the rest of $X \cup Y$, i.e. $(X \cup Y) \setminus (X \cap Y)$, if $\alpha > 1$ it is less influential. Both Marczewski-Steinhaus and Dice-Sørensen μ-indexes are special cases of the Symmetric Tversky μ-index, the former with $\alpha = 1$ and the latter with $\alpha = 0.5$.

The Tversky μ-index with $\alpha \neq \beta$ implies that $sim_{\mu T}^{\alpha,\beta}(X,Y) = sim_{\mu T}^{\alpha,\beta}(Y,X)$ will not hold for all $X = Y$, which is hard to justify and interpret in the setting of this thesis. It will be shown that the concept of optimal approximation proposed later does not work for the Tversky μ-index with $\alpha \neq \beta$.

Chapter 5

Optimal Approximations

It is generally easy to identify whether two finite groups are equal even though in general *equality* is undecidable[30]. In this thesis however, all sets of data are finite. If each group contains the exact same elements they are equal, and if they do not, they are not equal. But this is very limiting when comparing groups which are seldom *exactly* equal because it yields almost no information at all! The only derivable fact is that the groups are not identical according to a particular equality measure. Measure theory shows that two groups could be equal according to a certain type of measurement, without containing the same elements. It would be very helpful to be able to quantify how 'close to equal' (or distant) two sets are, so we take advantage of similarity measures discussed in the previous chapter.

For example, if we use weight as the type of size evaluation (μ = weight) then we would find a 5 kg bag of feathers to be equal to a 5 kg bowling ball. Obviously the sets are not identical, but from the perspective of weight, they are equal. Using another measurement such as size, or quantity of items in the set, the equality of the sets could change.

There are many methods of comparing sets of items. From the most simplistic comparisons which count the number of items in the intersection of sets, to the more complex comparisons that rely on weights or even item attributes. Comparing measurements of entire sets is not useful in general as the above example demonstrates. It neglects important and easily obtainable information, such as the measurement of the set of items common to both sets, and the measurement of items unique to each set.

Instead of comparing entire sets, we might also want to find an approximation of a single set, or if we are given an approximation, we might want to check how well it approximates the target set. It is easy to check two simple cases. If the two sets are exactly equal we assign a similarity of 100%, or restricted to the interval $[0, 1]$, a value of 1. If we determine that not a single item is common to both sets, this should correspond to a similarity of 0. Every other possibility lies between these two extremes.

In the field of rough sets, the concepts of upper and lower approximation which we explained in Section 3.2.1, are intrinsic. What has been neglected, and what we present here, is a method to determine the 'best' or 'optimal' approximation.

5.1 What is 'Optimal?'

Even though we know informally that the word *optimal* refers to being the 'best', it is not rigorous. So before properties of an optimal approximation can be discussed what it means to be 'optimal' must first be defined. Since optimal approximations will be explored in the context of rough sets, we can say that one rough set better approximates a target set, with respect to a given similarity measure, if its similarity

is greater than another. So it is possible to define what it means for one rough set to be better than all others at approximating some general set, by taking the largest similarity value. To ensure a valid context, let $\mathcal{AS} = (U, E)$ be a *Pawlak approximation space* where U is a finite and non-empty set, called the *universe*, and $E \subseteq U \times U$ is an *equivalence relation* on U.

A *general* definition of optimal approximation follows:

Definition 3. *For every set $X \subseteq U$, a definable set $\mathsf{O} \in \mathbb{D}$ is an **optimal approximation** of X (w.r.t. a given similarity measure sim that satisfies the axiom S2) if and only if:*

$$sim(X, \mathsf{O}) = \max_{\mathsf{A} \in \mathbb{D}}(sim(X, \mathsf{A}))$$

The set of all optimal approximations of X will be denoted by $\mathbf{Opt}_{sim}(X)$. □

Defining what it means for an approximation to be optimal makes it possible to begin comparing approximations, and each time two approximations of a set are compared, the less precise approximation is shown to not be optimal so it can be disregarded. A specific optimal approximation depends on the definition of the similarity measure sim. Of course if a different similarity measure is used, a different optimal approximation may be obtained even if the approximating and target set are the same in both cases, i.e. if $sim_1 \neq sim_2$ then clearly $\mathbf{Opt}_{sim_1}(X)$ might differ from $\mathbf{Opt}_{sim_2}(X)$ for some $X \subseteq U$.

It is also prudent to point out that definition 3 does not make much sense for the Tversky μ-index with $\alpha \neq \beta$. This is because assuming $X \neq \mathsf{A}$, then $sim^{\alpha,\beta}_{\mu T}(X, \mathsf{A}) > sim^{\alpha,\beta}_{\mu T}(\mathsf{A}, X) \iff \alpha < \beta$. In the rough sets approach there is no reason why the set $X \setminus \mathsf{A}$ should be treated differently than $\mathsf{A} \setminus X$. While similarities without the axiom S2 have some applications (for example to make a distinction between prototypes and

variants, c.f. [34]), they are not part of this thesis. Future work might examine how to use this index by fixing, or bounding the variables α and β. For more discussion on this issue, see Section 5.3.

Now to move toward an algorithm for determining the optimal approximation, first restrict the domain for the possible optimal rough approximating sets. From the axioms presented in Chapter 4, S4 and S5 imply that all optimal approximations reside between lower and upper approximations (inclusive), for all similarity indexes that satisfy them.

Proposition 2. *Assume that a similarity index $sim(...)$ satisfies the axioms S4–S5. Then, for every set $X \subseteq U$, and every $\mathsf{O} \in \mathbf{Opt}_{sim}(X)$:*

$$\underline{\mathbf{A}}(X) \subseteq \mathsf{O} \subseteq \overline{\mathbf{A}}(X)$$

Proof. Suppose that $\underline{\mathbf{A}}(X) \not\subseteq \mathsf{O}$, i.e. $C = \underline{\mathbf{A}}(X) \setminus \mathsf{O} \neq \emptyset$. Since $C \subseteq X$, then by axiom $S4$, $sim(\mathsf{O}, X) < sim(\mathsf{O} \cup C, X)$, so O must not be optimal. Now suppose that $\mathsf{O} \not\subseteq \overline{\mathbf{A}}(X)$, i.e. $C = \mathsf{O} \setminus \overline{\mathbf{A}}(X) \neq \emptyset$. By axiom S5, $sim(\mathsf{O} \setminus C, X) > sim(\mathsf{O}, X)$, so O must not be optimal again. □

We have to note here that the above result depends on the axiom S5 and its weakened version S5′ does not suffice, so Proposition 2 cannot be applied for Braun-Blanquet μ-index. Observe the following example.

Example 1. *Consider the universe of elements $U = \{a_1, a_2, b_1, b_2, b_3, c_1, c_2, c_3, c_4, c_5, c_6, c_7, d_1\}$ with equivalence classes $\mathfrak{Comp} = \{A, B, C, D\}$, where $A = \{a_1, a_2\}$, $B = \{b_1, b_2, b_3\}$, $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$, $D = \{d_1\}$ and select the set $X = \{a_1, b_1, b_2, c_1,$*

$c_2, c_3\}$. *Also assume that* $\mu(Y) = |Y|$ *for all* $Y \subseteq U$. *When we compare* X *to all definable sets using the Braun-Blanquet index, the maximum similarity value we obtain is* $\frac{1}{2}$. *We can get this by evaluating* $sim_{FS}(X, A \cup B)$, *or* $sim_{BB}(X, B \cup C)$, *or* $sim_{BB}(X, A \cup B \cup C)$. *By our definition for the definable set* $A \cup B \cup D$ *we get* $sim_{BB}(X, A \cup B \cup D) = \frac{1}{2}$ *which makes it an optimal approximation, but* $A \cup B \cup D \not\subseteq \overline{\mathbf{A}}(X) = A \cup B \cup C$.

The above demonstrates that optimal approximations using the Braun-Blanquet index cannot be bounded to be between the upper and lower approximations. So, the difficulty with this index is that there may be optimal approximations which cannot be determined except through an exhaustive search, which is not always viable.

The above proposition leads to the following trivial corollary.

Corollary 3. *For every* $X \subseteq U$, *and* $\mathsf{O} \in \mathbf{Opt}_{sim}(X)$,

$$sim(X, \underline{\mathbf{A}}(X)) \leq sim(X, \mathsf{O}), \text{ and } sim(X, \overline{\mathbf{A}}(X)) \leq sim(X, \mathsf{O}). \qquad \square$$

Since the similarity between X and O is at a maximum, the similarity must be smaller between X and either the upper or lower approximations. With the possible optimal approximations now bounded to be between the lower and upper approximations, we define this region, and the rough sets within it.

Definition 4. *Let* $X \subseteq U$, *and* $\mathsf{O} \in \mathbb{D}$. *We say that* O *is an* ***intermediate approximation*** *of* X, *if*

$$\underline{\mathbf{A}}(X) \subseteq \mathsf{O} \subseteq \overline{\mathbf{A}}(X)$$

The set of all intermediate approximations of X *will be denoted by* $\mathbf{IA}(X)$.

i.e. $\mathbf{IA}(X) = \{O | O \in \mathbb{D} \wedge \underline{\mathbf{A}}(X) \subseteq \mathsf{O} \subseteq \overline{\mathbf{A}}(X)\}$ $\square$

Note that the intermediate approximation is independent of a similarity index that is used to find an optimal approximation $\mathbf{Opt}_{sim}(X)$ since neither the lower nor upper approximations depend on a similarity index, but it is assumed that axioms S4 and S5 are satisfied.

For indexes which satisfy it, one of the consequences of Proposition 2 is that any optimal approximation of X, is the union of the lower approximation of X and some element $\mathsf{A} \in \mathbb{B}(X) \cup \{\emptyset\}$. It could also be represented as the upper approximation with some $\mathsf{A} \in \mathbb{B}(X) \cup \{\emptyset\}$ cut from it. For this and upcoming sections, recall the definitions of $\mathfrak{B}(X)$ and $\mathbb{B}(X)$ from Section 3.2.2 as they will be useful here.

From Proposition 2 we have:

Corollary 4. *For each set $X \subseteq U$,*

1. $\mathbf{Opt}_{sim}(X) \subseteq \mathbf{IA}_{sim}(X)$.
2. *If* $\mathsf{O} \in \mathbf{Opt}_{sim}(X)$ *then* $\exists \mathsf{A} \in \mathbb{B}(X) \cup \{\emptyset\}$ *such that* $\mathsf{O} = \underline{\mathbf{A}}(X) \cup \mathsf{A}$.
3. *If* $\mathsf{O} \in \mathbf{Opt}_{sim}(X)$ *then* $\exists \mathsf{B} \in \mathbb{B}(X) \cup \{\emptyset\}$ *such that* $\mathsf{O} = \overline{\mathbf{A}}(X) \setminus \mathsf{B}$. □

These are properties of optimal approximations. The first simply states that every optimal approximation is also an intermediate approximation. The last two explain formally what was mentioned above. Any optimal approximation must be the union of the lower approximation with some definable set which is in the Border (or is the empty set itself). Equivalently, it must also be possible to represent any optimal approximation by the upper approximation with some border set removed from it (or the empty set itself if the upper approximation is already optimal).

The notion of optimal approximation also introduces some structure to the current available field of similarity measures, as certain different similarity indexes may generate the same optimal approximations. Beginning more generally, we define a sort of equivalency among similarity measures. When we are looking for an optimal approximation we may not care what the result of the similarity evaluation yields, only which rough set satisfies the maximum similarity criteria, so in case an alternative measure will allow faster or easier computation, it can be substituted.

Definition 5. *We say that two similarity indexes sim_1 and sim_2 are **consistent** if for all sets A, B, C,*

$$sim_1(A,B) < sim_1(A,C) \iff sim_2(A,B) < sim_2(A,C). \qquad \square$$

Essentially, from the perspective of only comparing which gives a larger result, the indexes are equivalent. This clearly leads to the following result.

Corollary 5. *If sim_1 and sim_2 are consistent then for each $X \subseteq U$,*

1. $\mathbf{Opt}_{sim_1}(X) = \mathbf{Opt}_{sim_2}(X)$.

2. *sim_1 satisfies the axioms S4 and S5 if and only if sim_2 satisfies them.* $\square$

If two similarity measures are consistent, they will agree on the relative sizes of all pairs of similarity evaluations. This means that the set of optimal rough sets will be the same for all consistent measures. We can also conclude that measures which are consistent with each other, either all satisfy our monotonicity axioms, or none of them do. These concepts will allow us to extend results and algorithms designed for specific similarity indexes, to larger classes of consistent indexes.

First we will show that the Marczewski-Steinhaus μ-index and the Symmetric Tversky μ-index are consistent.

Proposition 3 (Consistency of Marczewski-Steinhaus and Sym. Tversky μ-indexes)**.** *For all A, B, C and $\alpha > 0$*

$$sim_{MS}(A,B) < sim_{MS}(A,C) \iff sim^{\alpha}_{\mu sT}(A,B) < sim^{\alpha}_{\mu sT}(A,C).$$

Proof. If $A = C$ then $sim_{MS}(A,C) = sim^{\alpha}_{\mu sT}(A,C) = 1$, so the equivalence holds. Assume $A \neq C$. Since $sim_{MS}(A,C) > 0$, then $A \cap C \neq \emptyset$. Moreover $A \setminus C \neq \emptyset$ or $C \setminus A \neq \emptyset$. Hence: $sim_{MS}(A,B) < sim_{MS}(A,C) \iff \frac{\mu(A\cap B)}{\mu(A\cup B)} < \frac{\mu(A\cap C)}{\mu(A\cup C)} \iff$
$\frac{\mu(A\cap B)}{\mu(A\cap B)+\mu(A\setminus B)+\mu(B\setminus A)} < \frac{\mu(A\cap C)}{\mu(A\cap C)+\mu(A\setminus C)+\mu(C\setminus A)} \iff$
$\mu(A\cap B)(\mu(A\setminus C)+\mu(C\setminus A)) < \mu(A\cap C)(\mu(A\setminus B)+\mu(B\setminus A)) \iff$
$\frac{\mu(A\cap B)}{\mu(A\cap C)} < \frac{\mu(A\setminus B)+\mu(B\setminus A)}{\mu(A\setminus C)+\mu(C\setminus A)} \iff \frac{\mu(A\cap B)}{\mu(A\cap C)} < \frac{\alpha\mu(A\setminus B)+\alpha\mu(B\setminus A)}{\alpha\mu(A\setminus C)+\alpha\mu(C\setminus A)} \iff$
$\frac{\mu(A\cap B)}{\mu(A\cap B)+\alpha\mu(A\setminus B)+\alpha\mu(B\setminus A)} < \frac{\mu(A\cap C)}{\mu(A\cap C)+\alpha\mu(A\setminus C)+\alpha\mu(C\setminus A)} \iff$
$sim^{\alpha}_{\mu sT}(A,B) < sim^{\alpha}_{\mu sT}(A,C).$ □

Since the Dice-Sørensen index is exactly the same as the symmetric Tversky μ-index with $\alpha = 0.5$, the above proposition immediately implies that the Dice-Sørensen and Marczewski-Steinhaus μ-indexes are consistent too. Thus we can begin to create a sort of equivalence class of similarity relations. With the following corollary the μMS, μDS, and μsT indexes are all consistent.

Corollary 6 (Consistency of Marczewski-Steinhaus and Dice-Sørensen μ-indexes)**.** *For all A, B, C,*

$$sim_{MS}(A,B) < sim_{MS}(A,C) \iff sim_{\mu DS}(A,B) < sim_{\mu DS}(A,C).$$

Proof. Since $sim_{\mu DS}(A,B) = sim^{0.5}_{\mu sT}(A,B)$. □

In general the Braun-Blanquet μ-index is *not consistent* with the Marczewski-Steinhaus index. To demonstrate this, consider the following example:

Example 2. *Consider a universe of 32 elements, with the following defined sets:* $A = \{a_1, a_2, a_3, a_4\}$, $B = \{a_1, a_2, a_3, a_5, ..., a_{21}\}$, $C = \{a_1, a_4, a_{22}, ..., a_{32}\}$, *and* μ *is cardinality, so* $\mu(X) = |X|$ *(so the Marczewski-Steinhaus* μ*-index is exactly the Jaccard index, and Fuzzy Sets* μ*-index is simply the Fuzzy Sets index). We can easily count the sizes of the sets:* $|A| = 4, |B| = 20, |C| = 13, |A \cap B| = 3$ *and* $|A \cap C| = 2$. *Hence* $sim_{MS}(A,B) = \frac{3}{21} = \frac{1}{7} > sim_{MS}(A,C) = \frac{2}{15}$, *while* $sim_{BB}(A,B) = \frac{3}{20} < sim_{BB}(A,C) = \frac{2}{13}$.

In this example, if we evaluate similarity using the MS index, B is more similar to A than C is to A, but if we use the BB index C is more similar. Thus, we clearly cannot substitute these indexes for one another.

The similarity indexes that do not satisfy the axiom S2 are not covered by the theory presented in this thesis, however for the sake of completeness, it is shown with the following example, that the Tversky index with $\alpha \neq \beta$ is not consistent with the Jaccard index.

Example 3. *Begin with a universe of 12 elements, and three sets.* $A = \{a_1, a_2, a_3, a_4\}$, $B = \{a_1, a_2, a_3, a_4, a_6, ..., a_{12}\}$, *and* $C = \{a_3, a_4, a_5\}$. *The sizes of each set are as follows:* $|A| = 4$, $|B| = 11$, $|C| = 3$, $|A \cap B| = 4$, $|A \cap C| = 2$, $|A \cup B| = 12$, *and* $|A \cup C| = 5$. *So in this case* $sim_J(A,B) = \frac{4}{11} < sim_J(A,C) = \frac{2}{5}$, *but for any* α *and*

β such that $\frac{\alpha}{\beta} > \frac{5}{4}$, we have $sim_T^{\alpha,\beta}(A,B) > sim_T^{\alpha,\beta}(A,C)$. For example for $\alpha = 1.5$ and $\beta = 1.0$ we have $sim_T^{\alpha,\beta}(A,B) = \frac{4}{11} > sim_T^{\alpha,\beta}(A,C) = \frac{1}{3}$. □

So in general, the Marczewski-Steinhaus μ-index and the Tversky μ-index are not consistent for $\alpha \neq \beta$.

Until now, the concept of optimal approximation has not been applied to any specific similarity measure. It was only assumed that whichever index is chosen as the function *sim*, it will satisfy the axioms $S1$–$S5$. However, to show more specific and detailed properties of optimal approximations, and in particular an efficient algorithm to find them, we choose a specific similarity measure for evaluation purposes. Due to Corollary 5(1), the results will hold for all other consistent similarity indexes.

5.2 Optimal Approximations using the Marczewski-Steinhaus Similarity Index

We chose to begin by using the Marczewski-Steinhaus μ-index because it is metrical, and has a natural and regular definition. This makes it perfect for discovering and proving mathematical results. However, we more often used the more specific Jaccard index to explore algebraic properties, and then verified that the results could be extended to the general case.

It is assumed going forward, that a Pawlak approximation space $\mathcal{AS} = (U, E)$ (i.e. U is finite) is available, as is a ***finite*** and ***null-free*** measure on U (defined element-wise) called $\mu : U \to \mathbb{R}$.

Definition 6. *For every $X, Y \subseteq U$, such that $X \setminus Y \neq \emptyset$, we define an index $\rho(X, Y)$, called the **similarity ratio**, or the ratio of common to distinct elements, as follows*

$$\rho(X, Y) = \frac{\mu(X \cap Y)}{\mu(X \setminus Y)}.$$

Note that $\rho(X, Y)$ is sound only if μ is finite and null-free.

By Proposition 1, the Marczewski-Steinhaus μ-index satisfies the axioms $S1$–$S5$, so the property specified by Proposition 2 and Definition 4 is satisfied.

With this index, we now have the necessary tools to define and prove a Lemma which will be central to this thesis. Given a non-definable target set X, and any definable set O which is known to be part of the upper approximation, but not part of the lower approximation, we can first infer that $\mathsf{O} \in \mathbf{IA}(X)$ and call it an intermediate approximation of X. Essentially this is the starting point. Now two situations may arise. If we are given an additional component $\mathbf{x} \in \mathfrak{B}(X)$(or set of components) from the border of X **which has no common element** with O, i.e. $\mathsf{O} \cap \mathbf{x} = \emptyset$ we can determine which whether O or $\mathsf{O} \cup \mathbf{x}$ is a better approximation of X, using part (1) of the lemma below. If instead of being disjoint, the additional (set of) component(s) is(are) **completely contained** within our intermediate set, instead use part (2) of the lemma to determine if a better approximation would be obtained using O or $\mathsf{O} \setminus \mathbf{x}$. So, beginning with any intermediate approximation, check if adding or subtracting any other definable set results in a better approximation.

Lemma 1. *Let* $X \subseteq U$, $\mathsf{O} \in \mathbf{IA}(X)$, $\mathsf{A}, \mathsf{B} \in \mathbb{B}(X)$, $\mathsf{A} \cap \mathsf{O} = \emptyset$, *and* $\mathsf{B} \subseteq \mathsf{O}$. *Then*

1. $sim_{MS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{MS}(X, \mathsf{O}) \iff \rho(\mathsf{A}, X) \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} = sim_{MS}(X, \mathsf{O})$

2. $sim_{MS}(X, \mathsf{O} \setminus \mathsf{B}) \leq sim_{MS}(X, \mathsf{O}) \iff \rho(\mathsf{B}, X) \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} = sim_{MS}(X, \mathsf{O})$

Proof. (1) To begin assign variables to make the algebra more clear. Let $\mu(X \cap \mathsf{O}) = n$, $\mu(X \cup \mathsf{O}) = m$, $\mu(\mathsf{A} \setminus X) = l$, and $\mu(\mathsf{A} \cap X) = k$. So the similarity ratio $\rho(\mathsf{A}, X) = \frac{k}{l}$. By Corollary 2(3) and the fact that μ is null-free, the values of n, m, l, k are all bigger than zero. It is known that $sim_{MS}(X, \mathsf{O}) = \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$ and $sim_{MS}(X, \mathsf{O} \cup \mathsf{A}) = \frac{\mu(X \cap (\mathsf{O} \cup \mathsf{A}))}{\mu(X \cup (\mathsf{O} \cup \mathsf{A}))}$. Because $\mathsf{A} \cap \mathsf{O} = \emptyset$, $\mu(X \cap (\mathsf{O} \cup \mathsf{A})) = \mu(X \cap \mathsf{O}) + \mu(X \cap \mathsf{A}) = n + k$ and $\mu(X \cup (\mathsf{O} \cup \mathsf{A})) = \mu(X \cup \mathsf{O}) + \mu(\mathsf{A} \setminus X) = m + l$. Now arrange a simple algebraic cancellation: $sim_{MS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{MS}(X, \mathsf{O}) \iff \frac{n+k}{m+l} \geq \frac{n}{m} \iff \frac{k}{l} \geq \frac{n}{m} \iff \rho(\mathsf{A}, X) \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} = sim_{MS}(X, \mathsf{O})$.

(2) Similar to the above, first assign $\mu(X \cap \mathsf{O}) = n$, $\mu(X \cup \mathsf{O}) = m$, $\mu(\mathsf{B} \setminus X) = l$, and $\mu(\mathsf{B} \cap X) = k$, so our similarity ratio is again $\rho(\mathsf{B}, X) = \frac{k}{l}$. By Corollary 2(3) and the definition of a null-free measure μ, the values of n, m, l, k are all bigger than zero. We have here $sim_{MS}(X, \mathsf{O}) = \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$ and $sim_{MS}(X, \mathsf{O} \setminus \mathsf{B}) = \frac{\mu(X \cap (\mathsf{O} \setminus \mathsf{B}))}{\mu(X \cup (\mathsf{O} \setminus \mathsf{B}))}$. Because $\mathsf{B} \subseteq \mathsf{O}$, $\mu(X \cap (\mathsf{O} \setminus \mathsf{B})) = \mu(X \cap \mathsf{O}) - \mu(X \cap \mathsf{B}) = n - k$ and $\mu(X \cup (\mathsf{O} \setminus \mathsf{B})) = \mu(X \cup \mathsf{O}) - \mu(\mathsf{B} \setminus X) = m - l$. Thus, $sim_{MS}(X, \mathsf{O} \setminus \mathsf{B}) \leq sim_{MS}(X, \mathsf{O}) \iff \frac{n-k}{m-l} \leq \frac{n}{m} \iff \frac{k}{l} \geq \frac{n}{m} \iff \rho(\mathsf{B}, X) \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} = sim_{MS}(X, \mathsf{O})$. □

Note that we cannot replace equations (1) and (2) of Lemma 1 by one equation, as the assumptions about A and B are entirely different. Moreover Lemma 1 does not hold if the measure μ is not null-free, as then the values of n, m, l, k from the proof of Lemma 1 are no longer bigger than zero. Since the above lemma is true for rough sets A and B, it clearly also holds for $\mathsf{A} = \mathbf{x} \in \mathfrak{B}(X)$ or $\mathsf{B} = \mathbf{x} \in \mathfrak{B}(X)$.

Intuitively, if we are trying to accurately approximate a target set and we are faced with a decision of whether to include a group, if more than half of the elements in the group under consideration are part of the target set then we would expect to need to include this group in the approximation. It is very convenient that we can define the corollary below to support this notion.

For a target set X if more than half of the elements of $\mathbf{x}$ also belong to X, (or equivalently, if more elements of $\mathbf{x}$ belong to X than do not) the rough set $\mathsf{O} \cup \mathbf{x}$ will approximate X better than O.

Corollary 7 ('Majority Rule')**.** *Let* $X \subseteq U$, $\mathsf{O} \in \mathbf{IA}(X)$, $\mathbf{x} \in \mathfrak{B}(X)$, *and* $\mathbf{x} \cap \mathsf{O} = \emptyset$. *Then:* $\mu(\mathbf{x} \cap X) \geq \mu(\mathbf{x} \setminus X) \iff \frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x})} \geq \frac{1}{2} \implies sim_{MS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{MS}(X, \mathsf{O})$.

Proof. Clearly $\mu(\mathbf{x} \cap X) \geq \mu(\mathbf{x} \setminus X) \iff \rho(\mathbf{x}, X) = \frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x} \setminus X)} \geq 1$. But $sim_{MS}(X, O) = \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} \leq 1$ because the Marczewski-Steinhaus μ-index satisfies axiom S1 as proven in Proposition 1. So, by using Lemma 1(1) reading right to left, conclude that $sim_{MS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{MS}(X, \mathsf{O})$. □

However, one must be cautious about even such simple claims because the converse of Corollary 7 does not hold. It may happen that $\frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x})} < \frac{1}{2}$, i.e. fewer than half of the elements in the set under consideration are part of the target set, but the rough set $\mathsf{O} \cup \mathbf{x}$ still approximates X better than O. Consider the following example:

Example 4. *Let* $\mathsf{O} = \{a_1, a_2, a_3, a_4, a_5\}$, $\mathbf{x} = \{b_1, b_2, b_3, b_4, b_5\}$, *and* $X = \{a_1, a_2, a_3, a_4, a_5, b_1, b_2, c_1\}$. *Assume that the measure* μ *is cardinality, i.e.* $\mu(A) = |A|$ *for all finite* A. *Then* $\frac{|\mathbf{x} \cap X|}{|\mathbf{x}|} = \frac{2}{5} = 0.4 < \frac{1}{2}$, *but* $sim_J(X, \mathsf{O} \cup \mathbf{x}) = \frac{7}{11} = 0.636 > sim_J(X, \mathsf{O}) = \frac{5}{8} = 0.6254$. □

From Proposition 2 notice that if $\mathsf{O} \in \mathbf{Opt}(X)$, then $\mathsf{O} = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_k$, for some $k \geq 0$, where each $\mathbf{x}_i \in \mathfrak{B}(X)$, $i = 1, \ldots, k$ and $k = 0$ corresponds to the case where $\mathsf{O} = \underline{\mathbf{A}}(X)$. Lemma 1 allows these $\mathbf{x}_i \in \mathfrak{B}(X)$ components to be explicitly defined.

Theorem 1. *For every $X \subseteq U$, the following two statements are equivalent:*

1. $\mathsf{O} \in \mathbf{Opt}(X)$
2. $\mathsf{O} \in \mathbf{IA}(X) \wedge \Big(\forall \mathbf{x} \in \mathfrak{B}(X).\, \mathbf{x} \subseteq \mathsf{O} \iff \rho(\mathbf{x}, X) = \frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x} \setminus X)} \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} = sim_{MS}(X, \mathsf{O})\Big)$.

Proof. (1)⇒(2) By Proposition 2, $\mathsf{O} \in \mathbf{IA}(X)$. Let $\mathbf{x} \in \mathfrak{B}(X)$ and $\mathbf{x} \subseteq \mathsf{O}$. To form a contradiction, suppose that $\frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x} \setminus X)} < \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$. Then by Lemma 1(2), $sim_{MS}(X, \mathsf{O} \setminus \mathbf{x}) > sim_{MS}(X, \mathsf{O})$, so O is not optimal. Next, assign $\mathbf{x} \in \mathfrak{B}(X)$ and $\mathbf{x} \cap \mathsf{O} = \emptyset$ and try $\frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x} \setminus X)} \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$. By Corollary 2(3), we have $\mathbf{x} \cap X \neq \emptyset$, so let $a \in \mathbf{x} \cap X$. Since $\mathbf{x} \cap \mathsf{O} = \emptyset$, then $a \in X \setminus \mathsf{O}$. Then by Proposition 1(1) and axiom $S4$, $sim_{MS}(X, \mathsf{O} \cup \{a\}) > sim_{MS}(X, \mathsf{O})$, so O is not optimal. Note that Lemma 1 gives only $sim_{MS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{MS}(X, \mathsf{O})$ which is not strong enough.

(2)⇒(1) Suppose O satisfies (2) but $\mathsf{O} \notin \mathbf{Opt}(X)$. Let $\mathsf{Q} \in \mathbf{Opt}(X)$. Hence, by the proof (1)⇒(2), Q satisfies (2). We have to consider two cases $\mathsf{Q} \setminus \mathsf{O} \neq \emptyset$ and $\mathsf{O} \setminus \mathsf{Q} \neq \emptyset$.

(*Case 1*) Let $\mathsf{Q} \setminus \mathsf{O} \neq \emptyset$ and let $\mathbf{y} \in \mathfrak{B}(X)$ be such that $\mathbf{y} \subseteq \mathsf{Q} \setminus \mathsf{O}$. Since Q satisfies (2), we have $\frac{\mu(\mathbf{y} \cap X)}{\mu(\mathbf{y} \setminus X)} \geq \frac{\mu(X \cap \mathsf{Q})}{\mu(X \cup \mathsf{Q})} = sim_{MS}(X, \mathsf{Q})$, and because $\mathsf{Q} \in \mathbf{Opt}(X)$, $sim_{MS}(X, \mathsf{Q}) \geq sim_{MS}(X, \mathsf{O})$. But by Lemma 1 this means that $\frac{\mu(\mathbf{y} \cap X)}{\mu(\mathbf{y} \setminus X)} \geq \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$. However O also satisfies (2) and $\mathbf{y} \in \mathfrak{B}(X)$, so by (2), $\mathbf{y} \subseteq \mathsf{O}$, a contradiction. Hence $\mathsf{Q} \setminus \mathsf{O} = \emptyset$.

(*Case 2*) Now we suppose that $\mathsf{O} \setminus \mathsf{Q} \neq \emptyset$ and let $\mathsf{O} \setminus \mathsf{Q} = \{\mathbf{y}_1, \ldots, \mathbf{y}_p\} \subseteq \mathfrak{B}(X)$. For clarity, we assign variables, so let $\mu(X \cap \mathsf{O}) = n$, $\mu(X \cup \mathsf{O}) = m$, and $\mu(\mathbf{y}_i \setminus X) = l_i$,

$\mu(\mathbf{y}_i \cap X) = k_i$, for $i = 1, \ldots, p$. Since O satisfies (2), for each $i = 1, \ldots, p$, we have $\frac{\mu(\mathbf{y}_i \cap X)}{\mu(\mathbf{y}_i \setminus X)} \geq \frac{|X \cap \mathsf{O}|}{|X \cup \mathsf{O}|}$, or equivalently $\frac{k_i}{l_i} \geq \frac{n}{m}$. Since it is valid for all i we can get

$$(k_1 + \ldots + k_p)m \geq (l_1 + \ldots + l_p)n.$$

On the other hand, since Q is optimal and we suppose O is not, $sim_{MS}(X, \mathsf{Q}) > sim_{MS}(X, \mathsf{O})$. From case 1, we know that there are no elements in Q which are not in O ($\mathsf{Q} \setminus \mathsf{O} = \emptyset$) so $sim_{MS}(X, \mathsf{Q}) = sim_{MS}(X, \mathsf{O} \setminus (\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p))$, then if we assign $\mathsf{B} = \{\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p\}$ by Lemma 1(2) we get $\frac{\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \cap X)}{\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \setminus X)} < \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})}$. Because $\mathbf{y}_i$ are components, we have $\mathbf{y}_i \cap \mathbf{y}_j = \emptyset$ when $i \neq j$. Thus $\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \cap X) = \mu(\mathbf{y}_1 \cap X) + \ldots + \mu(\mathbf{y}_p \cap X) = k_1 + \ldots + k_p$, and $\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \setminus X) = \mu(\mathbf{y}_1 \setminus X) + \ldots + \mu(\mathbf{y}_p \setminus X) = l_1 + \ldots + l_p$. This means that we have

$$\frac{\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \cap X)}{\mu((\mathbf{y}_1 \cup \ldots \cup \mathbf{y}_p) \setminus X)} < \frac{\mu(X \cap \mathsf{O})}{\mu(X \cup \mathsf{O})} \iff \frac{k_1 + \ldots + k_p}{l_1 + \ldots + l_p} < \frac{n}{m},$$

and since we have the left side as valid above, then cross multiplying the right side yields

$$(k_1 + \ldots + k_p)m < (l_1 + \ldots + l_p)n,$$

which is a contradiction, i.e. $\mathsf{O} \setminus \mathsf{Q} = \emptyset$. Thus, $\mathsf{Q} \setminus \mathsf{O} = \emptyset$ and $\mathsf{O} \setminus \mathsf{Q} = \emptyset$, i.e., $\mathsf{Q} = \mathsf{O}$, so $\mathsf{O} \in \mathbf{Opt}(X)$. □

Theorem 1 gives the necessary and sufficient conditions for optimal approximations of X (with respect to the Marczewski-Steinhaus index and a given measure $\mu : U \to \mathbb{R}$) in terms of the elements of $\mathfrak{B}(X)$. We will use it to build an efficient algorithm for finding optimal approximations.

By Theorem 1, the value of $\rho(\mathbf{x}, X)$ will indicate if $\mathbf{x} \in \mathfrak{B}(X)$ is a part of an optimal approximation of X, or not. Since $\mathfrak{B}(X)$ is finite, its elements can be enumerated by natural numbers $1, \ldots, |\mathfrak{B}(X)|$.

- Assume that $r = |\mathfrak{B}(X)|$, $\mathfrak{B}(X) = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ and also

$$i \leq j \iff \rho(\mathbf{x}_i, X) \geq \rho(\mathbf{x}_j, X).$$

In other words, sort $\mathfrak{B}(X)$ by decreasing values of $\rho(\mathbf{x}, X)$. This sorting will be used to build a special sequence of intermediate approximations.

Let $\mathsf{O}_0, \mathsf{O}_1, \ldots, \mathsf{O}_r \in \mathbf{IA}(X)$ be the sequence of intermediate approximations of X defined for $i = 0, \ldots, r-1$ as follows: $\mathsf{O}_0 = \underline{\mathbf{A}}(X)$ and

$$\mathsf{O}_{i+1} = \begin{cases} \mathsf{O}_i \cup \mathbf{x}_{i+1} & \text{if } sim_{MS}(X, \mathsf{O}_i \cup \mathbf{x}_{i+1}) \geq sim_{MS}(X, \mathsf{O}_i) \\ \mathsf{O}_i & \text{otherwise.} \end{cases}$$

Note that usually $\mathsf{O}_r \neq \overline{\mathbf{A}}(X) = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_r$, since $\mathsf{O}_r = \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_r$, only if $sim_{MS}(X, \mathsf{O}_i \cup \mathbf{x}_{i+1}) \geq sim_{MS}(X, \mathsf{O}_i)$ for all $i = 0, \ldots, r-1$, or equivalently, if $\mathsf{O}_i = \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_i$ for $i = 1, \ldots, r$.

We claim that at least one of these O_i's is an optimal approximation. The following technical result is needed to prove this claim.

Lemma 2. *Let $k_1, \ldots, k_n$ and $l_1, \ldots, l_n$ be positive numbers such that $\frac{k_1}{l_1} \geq \frac{k_i}{l_i}$ for $i = 1, \ldots, n$. Then $\frac{k_1}{l_1} \geq \frac{k_1 + \ldots + k_n}{l_1 + \ldots + l_n}$.*

Proof. $\frac{k_1}{l_1} \geq \frac{k_i}{l_i}$ implies $k_1 l_i \geq k_i l_1$ for $i = 1, \ldots, n$. Hence $k_1 l_1 + k_1 l_2 + \ldots + k_1 l_n \geq$

$k_1 l_1 + k_2 l_1 + \ldots + k_n l_1 \iff \frac{k_1}{l_1} \geq \frac{k_1+\ldots+k_n}{l_1+\ldots+l_n}$, which ends the proof. □

The essential properties of the sequence $\mathsf{O}_0, \mathsf{O}_1, \ldots, \mathsf{O}_r$ are provided by the following theorem.

Theorem 2. *For every $X \subseteq U$, we set $r = |\mathfrak{B}(X)|$, and we have*

1. $sim_{MS}(X, \mathsf{O}_{i+1}) \geq sim_{MS}(X, \mathsf{O}_i)$, *for* $i = 0, \ldots, r-1$.
2. *If* $\rho(\mathbf{x}_1, X) \leq sim_{MS}(X, \underline{\mathbf{A}}(X))$ *then* $\underline{\mathbf{A}}(X) \in \mathbf{Opt}(X)$.
3. *If* $\rho(\mathbf{x}_r, X) \geq sim_{MS}(X, \overline{\mathbf{A}}(X))$ *then* $\overline{\mathbf{A}}(X) \in \mathbf{Opt}(X)$.
4. *If* $sim_{MS}(X, \mathsf{O}_p) \leq \rho(\mathbf{x}_p, X)$ *and* $sim_{MS}(X, \mathsf{O}_{p+1}) > \rho(\mathbf{x}_{p+1}, X)$, *then* $\mathsf{O}_p \in \mathbf{Opt}(X)$, *for* $p = 1, \ldots, r-1$.
5. *If* $sim_{MS}(X, \mathsf{O}_r) \leq \rho(\mathbf{x}_r, X)$ *then* $\mathsf{O}_r = \overline{\mathbf{A}}(X) \in \mathbf{Opt}(X)$.
6. *If* $\mathsf{O}_p \in \mathbf{Opt}(X)$, *then* $\mathsf{O}_i = \mathsf{O}_p$ *for all* $i = p+1, \ldots, r$. *In particular* $\mathsf{O}_r \in \mathbf{Opt}(X)$.
7. $\mathsf{O} \in \mathbf{Opt}(X) \implies \mathsf{O} \subseteq \mathsf{O}_p$, *where p is the smallest one from (6).*

Proof.(1) Immediately from Lemma 1 and the definition of the sequence $\mathsf{O}_0, \ldots, \mathsf{O}_r$.

(2) From Proposition 2 we have that if $\mathsf{O} \in \mathbf{Opt}(X)$, then $\mathsf{O} = \underline{\mathbf{A}}(X) \cup \mathbf{x}_{i_1} \cup \ldots \cup \mathbf{x}_{i_s}$ for some $i_j \in \{1, \ldots, r\}$. Since $\rho(\mathbf{x}_1, X) \geq \rho(\mathbf{x}_{i_j}, X)$ for $j = 1, \ldots, s$, by Lemma 2, $\rho(\mathbf{x}_1, X) \geq \frac{\mu((\mathbf{x}_{i_1} \cup \ldots \cup \mathbf{x}_{i_s}) \cap X)}{\mu((\mathbf{x}_{i_1} \cup \ldots \cup \mathbf{x}_{i_s}) \setminus X)}$. Hence $sim_{MS}(X, \underline{\mathbf{A}}(X)) \geq \frac{\mu((\mathbf{x}_{i_1} \cup \ldots \cup \mathbf{x}_{i_s}) \cap X)}{\mu((\mathbf{x}_{i_1} \cup \ldots \cup \mathbf{x}_{i_s}) \setminus X)}$, so by Lemma 1, $sim_{MS}(X, \underline{\mathbf{A}}(X)) \geq sim_{MS}(X, \mathsf{O})$, which means $\underline{\mathbf{A}}(X) \in \mathbf{Opt}(X)$.

(3) Note that $\rho(\mathbf{x}_r, X) \geq sim_{MS}(X, \overline{\mathbf{A}}(X))$ implies $\rho(\mathbf{x}_i, X) \geq sim_{MS}(X, \overline{\mathbf{A}}(X))$ for all $i = 1, \ldots, r$. Hence by Theorem 1, $\overline{\mathbf{A}}(X) \in \mathbf{Opt}(X)$.

(4) We have $sim_{MS}(X, \mathsf{O}_0) \leq sim_{MS}(X, \mathsf{O}_1) \leq \ldots \leq sim_{MS}(X, \mathsf{O}_r)$ and $\rho(\mathbf{x}_1, X) \geq \rho(\mathbf{x}_2, X) \geq \ldots \geq \rho(\mathbf{x}_r, X)$. Hence the property $sim_{MS}(X, \mathsf{O}_p) \leq \rho(\mathbf{x}_p, X)$ implies $\mathsf{O}_i = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_i$ for all $i = 1, \ldots, p$. Adding the property $sim_{MS}(X, \mathsf{O}_{p+1}) > \rho(\mathbf{x}_{p+1}, X)$ implies $\mathsf{O}_i = \mathsf{O}_p$ for all $i = p+1, \ldots, r$, which meant that $\mathsf{O}_p = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_p$ satisfies (2) of Theorem 1. Hence $\mathsf{O}_p \in \mathbf{Opt}(X)$.

(5) Again the property $sim_{MS}(X, \mathsf{O}_r) \leq \rho(\mathbf{x}_r, X)$ implies $\mathsf{O}_r = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_r$, so $\mathsf{O}_r = \overline{\mathbf{A}}(X)$. Additionally $\mathsf{O}_r = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_r$ satisfies (2) of Theorem 1, so $\mathsf{O}_r \in \mathbf{Opt}(X)$.

(6) From the proofs of (4) and (5).

(7) We have to show that if $\mathsf{O} = \underline{\mathbf{A}}(X) \cup \mathsf{A} \in \mathbf{Opt}(X)$, where $\mathsf{A} \in \mathbb{B}(X)$, then $\mathsf{A} \subseteq \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_p$. Suppose $\mathbf{x}_j \subseteq \mathsf{A}$ and $j > p$. Then $\rho(\mathbf{x}_j, X) < sim_{MS}(X, \mathsf{O}_p) = sim_{MS}(X, \mathsf{O})$, so O does not satisfy (2) of Theorem 1. Hence $\mathsf{A} \subseteq \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_p$. □

Point (1) of Theorem 2 states that O_{i+1} is a better (or equal) approximation of X than O_i, (2) and (3) characterize the cases when either $\underline{\mathbf{A}}(X)$ or $\overline{\mathbf{A}}(X)$ are optimal approximations, while (4) shows conditions when some O_p is an optimal approximation. Point (6) states that once O_p is found to be optimal, calculations can be stopped as the remaining O_{p+i} are the same as O_p, and the last point, (6) indicates that O_p is the greatest optimal approximation.

Algorithm 1 (Finding the Greatest Optimal Approximation)**.** *Let $X \subseteq U$.*

1. *Construct $\underline{\mathbf{A}}(X)$, $\overline{\mathbf{A}}(X)$, and $\mathfrak{B}(X)$. Assume $r = |\mathfrak{B}(X)|$.*
2. *For each $\mathbf{x} \in \mathfrak{B}(X)$, calculate $\rho(\mathbf{x}, X) = \frac{\mu(\mathbf{x} \cap X)}{\mu(\mathbf{x} \setminus X)}$.*
3. *Order $\rho(\mathbf{x}, X)$ in decreasing order and number the elements of $\mathfrak{B}(X)$ by this*

order, so $\mathfrak{B}(X) = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ *and* $i \leq j \iff \rho(\mathbf{x}_i, X) \geq \rho(\mathbf{x}_j, X)$.

4. *If* $\rho(\mathbf{x}_1, X) \leq sim_{MS}(X, \underline{\mathbf{A}}(X))$ *then* $\mathsf{O} = \underline{\mathbf{A}}(X)$.

5. *If* $\rho(\mathbf{x}_r, X) \geq sim_{MS}(X, \overline{\mathbf{A}}(X))$ *then* $\mathsf{O} = \overline{\mathbf{A}}(X)$.

6. *If neither (4) nor (5) is applied, calculate* O_p, *starting from* $p = 1$ *and increasing* p *by* 1, *until* $sim_{MS}(X, \mathsf{O}_{p+1}) > \rho(\mathbf{x}_{p+1}, X)$. *If* $sim_{MS}(X, \mathsf{O}_{p+1}) > \rho(\mathbf{x}_{p+1}, X)$ *holds, set* $\mathsf{O} = \mathsf{O}_p$. □

Note that the biggest p in (6) of the above algorithm is $r - 1$. However, due to Theorem 2, step (5) of the above algorithm covers the case $\mathsf{O} = \mathsf{O}_r$. Theorem 2 also guarantees that one of (4), (5), or (6) with $1 \leq p < r$ always holds. The case (4) of the above algorithm means that the optimal approximation O satisfies $\mathsf{O} = \mathsf{O}_0$, the case (5) corresponds to $\mathsf{O} = \mathsf{O}_r = \overline{\mathbf{A}}(X) = \underline{\mathbf{A}}(X) \cup \mathbf{x}_1 \cup \ldots \cup \mathbf{x}_r$, and the case (6) corresponds to all other cases.

From Theorem 2, O must be the greatest optimal approximation, i.e. $\mathsf{O} \in \mathbf{Opt}(X)$, and for all $\mathsf{O}' \in \mathbf{Opt}(X)$, $\mathsf{O}' \subseteq \mathsf{O}$. It is also known that $sim_{MS}(X, \mathsf{O}') = sim_{MS}(X, \mathsf{O})$

This *greedy* algorithm (because of the choice of $\rho(\mathbf{x}, X)$, c.f. [15]) has a complexity of $C_1 + C_2 + \mathrm{O}(r\log r)$, where C_1 is the complexity of constructing $\underline{\mathbf{A}}(X)$, $\overline{\mathbf{A}}(X)$, and $\mathfrak{B}(X)$; while C_2 is the complexity to assign $\mu(x)$ for each $x \in U$. Algorithms with $C = \mathrm{O}(|U|^2)$ can be found for example in [28], and clearly $C_2 = \mathrm{O}(|U|)$.

The most crucial line of the algorithm, line (6), runs in $\mathrm{O}(r)$, but line (3) involves sorting which has complexity $\mathrm{O}(r log r)$. Since $r < |U|$, the total complexity is $\mathrm{O}(|U|^2)$.

Algorithm 1 gives us the greatest optimal approximation O, however the whole

set $\mathbf{Opt}(X)$ can easily be derived from O just by subtracting appropriate elements of $\mathfrak{B}(X)$.

Note that because of Corollary 5(1), Algorithm 1 is also effective for *any* similarity measure *sim* that is *consistent* with the Marczewski-Steinhaus μ-index sim_{MS}, for any finite and null-free measure μ. In particular, by Proposition 3 and Corollary 6 we can use it for Symmetric Tversky μ-index $sim_{\mu T}$, Dice-Sørensen μ-index $sim_{\mu DS}$, and of course the classical Jaccard index sim_J.

Algorithm 1 requires the measure μ to be finite and null-free. The assumption of finiteness of μ is essential (c.f. [18]), but null-freeness is merely technical. If μ is not null-free, we can use the algorithm presented below.

Algorithm 2 (μ is not null-free)**.** *Let $\mathcal{AS} = (U, E)$ be a Pawlak approximation space, $\mu : U \to \mathbb{R}$ be a given measure that is finite but not null-free, and $X \subseteq U$.*

1. *Define $U' = U \setminus \{x \mid \mu(x) = 0\}$, $E' = E \cap (U' \times U')$, $X' = X \setminus \{x \mid \mu(x) = 0\}$ and $\mathcal{AS}' = (U', E')$.*
2. *Apply Algorithm 1 for X' and $\mathcal{AS}'$. Let O' be the outcome of this application.*
3. *Pick any $\mathsf{O} \in \mathbf{IA}(X)$ such that $\mathsf{O}' \subseteq \mathsf{O}$.*

Since $\mathsf{O}' \in \mathbf{Opt}(X')$ then $\mathsf{O} \in \mathbf{Opt}(X)$. Moreover $\mu(X \setminus X') = \mu(\mathsf{O} \setminus \mathsf{O}') = 0$. □

5.3 Asymmetry

It is appropriate here to pause to discuss the assumption of symmetry. The most fundamental reason for this restriction is that without it, one could not define what

it means for a rough set to be the *optimal* approximation. Although, it might be possible to distinguish somehow between the rough set most similar to a target set, and the rough set to which a target set is most similar (the seemingly trivial difference only being directionality) it has traditionally been assumed that these are equal, and so $sim(X, Y) = sim(Y, X)$. Without this assumption, the triangle inequality axiom is not valid, and without the axiom, it is not possible to determine if adding a particular equivalence class to an approximation will make the approximation better or worse. In addition, the Jaccard, Dice-Sørensen, Marczewski-Steinhaus, and Braun-Blanquet indexes are all symmetric, and the Tversky index is symmetric if $\alpha = \beta$. So, only with the Tversky index would it be possible to investigate non-symmetric properties, although we admit possible causes for such. One idea might be to define α and β to be proportional to the size of the sets, or their relative difference in size. If enough data are available, one could use other features than those which partitioned the universe to define the values.

As mentioned in the background review (see Sections 2.5 and 2.6), Tversky took issue with the assumption that similarity is symmetric. In *Features of Similarity*[34] he explained that geometric models representing objects as points in coordinate space had been the focus of theoretical analysis of similarity relations. From the geometric approach, a metric distance function is only logical. This function assigns a non-negative number representing distance, to every pair of points, and satisfies the axioms of minimality, symmetry, and the triangle inequality (see Sec. 2.5 for more). This could be one underlying reason that symmetry is so often assumed. Another speculative reason could be the perceived relative usefulness and need (or lack thereof) for the study of asymmetric properties. Nonetheless, Tversky wanted to illustrate to

the scientific community the ubiquity of asymmetric judgments of similarity. "[T]he present paper," he wrote, "provides empirical evidence for asymmetric similarities and argues that similarity should not be treated as a symmetric relation." [34]

The Tversky Index, originally proposed as the *ratio model* by Tversky in [34], generalizes other similarity indexes (as was illustrated in the introduction to them) and still allows for asymmetry when $\alpha \neq \beta$. Tversky conducted studies which surveyed students' opinions regarding distances between certain entities, including countries, figures, and letters, where he showed that asymmetry was common in judgments of similarity.

Tversky stressed that "this asymmetry in the *choice* of similarity statements is associated with asymmetry in *judgments* of similarity." [34] He also noted that this asymmetry is highlighted in the context of simile and metaphor. These situations are opinion, and language based. According to this author, the most useful applications of optimal rough sets lie in the field of pattern recognition. Areas which involve approximation due to either visual object recognition, or categorical data where a 'best' group is sought.

When Lee Dice produced the coincidence index[5], he was looking for a measurement of how associated two species were, and began with the association index, which was certainly not symmetric since it divided the number of samples where two species occurred together by the number where one occurred alone. Dice rationalized that there are many cases where one species is more dependent on another than the reverse, though one might counter that this is not a valid application for measuring similarity or distance.

5.4 Application

Due to the high level nature of Algorithm 1 and the field of rough sets in general, the algorithm would have almost endless possible applications. Whenever data are collected, and can be partitioned into classes of objects which can be distinguished by certain features, our algorithm could potentially be applied.

For example, in the fields of medicine and genomics, there are numerous reasons to compare DNA sequences, symptoms, or other attributes of patients. Determining the best match for transplants, or ideal candidates to derive virus treatments, or to isolate the most important patients to screen for diseases are all examples of possible applications. While the algorithm presented here may not isolate a single donor or recipient, it could quickly narrow down the scope of people to search through using traditional means. In the area of medicine, often each patient corresponds to a list of attributes regarding their current health and medical history. If a doctor or researcher wishes to isolate the best group to administer a trial drug, or figure out the patients who have similar or intersection histories which led them to similar conditions, they can develop their own algorithms or conduct a search manually. The field of rough sets using the upper and lower approximations can narrow down selection criteria to those patients that should be ruled out, and those that should certainly be included because all criteria are satisfied. However, the notion of an optimal rough set would allow a quick isolation down to the most relevant set of classes of patients, or the group of patients which best correspond to desired criteria. It may still be required for the group to be checked for how close each individual fits within the class but isolating a small group from a much larger population is often quite valuable when working with extremely large sets of data.

While it is true that certain human assumptions must be made, such as which features to use to determine the class of each item, these can be chosen to guide our algorithm into providing relevant results. The unfortunate part of trying to use Algorithm 2 in a brute force sense, is that it relies highly on the equivalence classes formed to partition the universe, as well as the evaluation of size, which are chosen by the researcher using the algorithm. This seems like a compelling opportunity to use appropriate machine learning techniques (e.g.neural networks, decision trees, support vector machines, etc.) or multiple size evaluation measures to derive an optimal partition of the universe. The machine learning could also be applied to multiple size evaluation mechanisms to determine which partition results in the most similar optimal approximation to the target set. This suggests the possibility that similarity relations should be written, for example, sim_J^E where J is the similarity index, and E is the equivalence relation used to partition the universe.

5.5 The Case of Fuzzy Set/Braun-Blanquet Index

The *Fuzzy set index* [26], defined as $sim_{FS}(X,Y) = \frac{|X \cap Y|}{\max(|X|,|Y|)}$, is inconsistent with the Jaccard index and our algorithm has only limited use with it because its natural (naive) extension *does not work* in general. One may be tempted just to use Algorithm 1 with sim_{MS} replaced by sim_{FS}, especially that it may actually work for 'regular' random cases. However it does not always work as the equivalence of Lemma 1, namely:

- Let $X \subseteq U$, $\mathsf{O} \in \mathbf{IA}(X)$, $\mathsf{A}, \mathsf{B} \in \mathbb{B}(X)$, $\mathsf{A} \cap \mathsf{O} = \emptyset$, and $\mathsf{B} \subseteq \mathsf{O}$. Then

  1. $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{FS}(X, \mathsf{O}) \iff \frac{|\mathsf{A} \cap X|}{|\mathsf{A} \setminus X|} \geq sim_{FS}(X, \mathsf{O})$

$$2.\ sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) \leq sim_{FS}(X, \mathsf{O}) \iff \frac{|\mathsf{B} \cap X|}{|\mathsf{B} \setminus X|} \geq sim_{FS}(X, \mathsf{O})$$

is *not true* in general. Some additional assumptions are required.

Consider the following two examples.

Example 5. *We begin with a universe* $U = \{a_1, a_2, b_1, b_2, \ldots, b_9, c_1, c_2, \ldots, c_{11}\}$*, three equivalence classes covering* U*,* $\mathsf{A}_1 = \{a_1, a_2\}$*,* $\mathsf{A}_2 = \{b_1, b_2, \ldots, b_9\}$*,* $\mathsf{A}_3 = \{c_1, c_2, \ldots, c_{11}\}$*, and the target set we wish to approximate is* $X = \{a_1, a_2, b_1, b_2, c_1, c_2\}$*. The lower and upper approximations are* $\underline{\mathbf{A}}(X) = \mathsf{A}_1 = \{a_1, a_2\}$ *and* $\overline{\mathbf{A}}(X) = \mathsf{A}_1 \cup \mathsf{A}_2 \cup \mathsf{A}_3 = U$ *respectively. One may check by inspection that* $\mathbf{Opt}_{sim_J}(X) = \underline{\mathbf{A}}(X) = \mathsf{A}_1$*, while* $\mathbf{Opt}_{sim_{FS}}(X) = \mathsf{A}_1 \cup \mathsf{A}_2$*. When applying Algorithm 1 with* sim_J *replaced by* sim_{FS} *we will get* A_1 *as an optimal approximation. The reason is that* $sim_{FS}(X, \mathsf{A}_1 \cup \mathsf{A}_2) = \frac{|X \cap (\mathsf{A}_1 \cup \mathsf{A}_2)|}{\max(|X|, |\mathsf{A}_1 \cup \mathsf{A}_2|)} = \frac{4}{11} = 0.364 > sim_{FS}(X, \mathsf{A}_1) = \frac{|X \cap \mathsf{A}_1|}{\max(|X|, |\mathsf{A}_1|)} = \frac{2}{6} = 0.333$*, but* $\frac{|\mathsf{A}_2 \cap X|}{|\mathsf{A}_2 \setminus X|} = \frac{2}{7} = 0.286 < sim_{FS}(X, \mathsf{A}_1) = 0.333$*, so the equivalent of Lemma 1 is not satisfied. Hence the first step of a modified Algorithm 1 would be faulty. Note also that* $sim_{FS}(X, \mathsf{A}_1 \cup \mathsf{A}_2) > sim_{FS}(X, \mathsf{A}_1)$ *while* $sim_J(X, \mathsf{A}_1 \cup \mathsf{A}_2) = \frac{4}{11} = 0.364 < sim_J(X, \mathsf{A}_1) = \frac{2}{5} = 0.4$*, so this illustrates the weak inconsistency between the Jaccard and Fuzzy Sets indexes.* □

Example 6. *Consider a universe* $U = \{a_1, a_2, b_1, b_2, \ldots, b_6, c_1, c_2, \ldots, c_{30}\}$*, three equivalence classes covering* U*,* $\mathsf{A}_1 = \{a_1, a_2\}$*,* $\mathsf{A}_2 = \{b_1, b_2, \ldots, b_6\}$ *and* $\mathsf{A}_3 = \{c_1, c_2, \ldots, c_{25}\}$*, and the set* $X = \{a_1, a_2, b_1, c_1, c_2, \ldots, c_5\}$*. We have* $\underline{\mathbf{A}}(X) = \mathsf{A}_1 = \{a_1, a_2\}$*,* $\overline{\mathbf{A}}(X) = \mathsf{A}_1 \cup \mathsf{A}_2 \cup \mathsf{A}_3 = U$*. One may check by inspection that* $\mathbf{Opt}_{sim_J}(X) = \underline{\mathbf{A}}(X) = \mathsf{A}_1$*, while* $\mathbf{Opt}_{sim_{FS}}(X) = \mathsf{A}_1 \cup \mathsf{A}_2$*. When applying Algorithm 1 with* sim_J *replaced by* sim_{FS} *we will get* A_1 *as an optimal approximation. The reason is that* $sim_{FS}(X, \mathsf{A}_1 \cup \mathsf{A}_2) = \frac{|X \cap (\mathsf{A}_1 \cup \mathsf{A}_2)|}{\max(|X|, |\mathsf{A}_1 \cup \mathsf{A}_2|)} = \frac{3}{9} = 0.333 > sim_{FS}(X, \mathsf{A}_1) = \frac{|X \cap \mathsf{A}_1|}{\max(|X|, |\mathsf{A}_1|)} = \frac{2}{9} = 0.222$*, but* $\frac{|\mathsf{A}_2 \cap X|}{|\mathsf{A}_2 \setminus X|} = \frac{1}{5} = 0.2 < sim_{FS}(X, \mathsf{A}_1) = 0.222$*, so the equivalent of Lemma 1*

is not satisfied either. Hence the first step of a modified Algorithm 1 would be faulty in this case as well. Note also that $sim_{FS}(X, \mathsf{A}_1 \cup \mathsf{A}_2) > sim_{FS}(X, \mathsf{A}_1)$ *while* $sim_J(X, \mathsf{A}_1 \cup \mathsf{A}_2) = \frac{3}{14} = 0.214 < sim_J(X, \mathsf{A}_1) = \frac{2}{9} = 0.222$*, so this is another example of weak inconsistency between Jaccard and Fuzzy Sets indexes.* □

For the Fuzzy Set index we can show the following result.

Lemma 3. *Let* $X \subseteq U$*,* $\mathsf{O} \in \mathbf{IA}(X)$*,* $\mathsf{A}, \mathsf{B} \in \mathbb{B}(X)$*,* $\mathsf{A} \cap \mathsf{O} = \emptyset$*, and* $\mathsf{B} \subseteq \mathsf{O}$*.*

1. **If** $|X| < |\mathsf{O}|$ **or** $|\mathsf{O} \cup (\mathsf{A} \cap X)| > |X| > |\mathsf{O}|$**, then**
   $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{FS}(X, \mathsf{O}) \iff \frac{|\mathsf{A} \cap X|}{|\mathsf{A} \setminus X|} \geq sim_{FS}(X, \mathsf{O})$
2. $sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) \leq sim_{FS}(X, \mathsf{O}) \iff \frac{|\mathsf{B} \cap X|}{|\mathsf{B} \setminus X|} \geq sim_{FS}(X, \mathsf{O})$

Proof. (1) First define $|X| = r$, $|\mathsf{O}| = s$, $|X \cap \mathsf{O}| = n$, $|X \cup \mathsf{O}| = m$, $|\mathsf{A} \setminus X| = l$ and $|\mathsf{A} \cap X| = k$. Clearly r, s are bigger than zero, and additionally, by Corollary 2(3), n, m, l, k are all bigger than zero. Also note that $|A| = l + k$, and, since $\mathsf{A} \cap \mathsf{O} = \emptyset$, $|\mathsf{O} \cup \mathsf{A}| = |\mathsf{O}| + |\mathsf{A}| = s + l + k$ and $|X \cap (\mathsf{O} \cup \mathsf{A})| = |X \cap \mathsf{O}| + |X \cap \mathsf{A}| = n + k$.
Case: $|X| < |\mathsf{O}|$. Here we have $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) = \frac{|X \cap (\mathsf{O} \cup \mathsf{A})|}{\max(|X|, |\mathsf{O} \cup \mathsf{A}|)} = \frac{n+k}{s+l+k}$ and $sim_{FS}(X, \mathsf{O}) = \frac{|X \cap \mathsf{O}|}{\max(|X|, |\mathsf{O}|)} = \frac{n}{s}$, so $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{FS}(X, \mathsf{O})$ means $\frac{n+k}{s+l+k} \geq \frac{n}{s}$. Simple algebra yields $\frac{n+k}{s+l+k} \geq \frac{n}{s} \iff ks \geq nl + kn$. Since all values are positive, if ks is greater than the sum, it must be greater than each term, so $ks \geq nl + kn \iff ks \geq nl \iff \frac{k}{l} \geq \frac{n}{s} = sim_{FS}(X, \mathsf{O})$ and thus we have proven the first case.
Case: $|\mathsf{O}| < |X| < |\mathsf{O} \cup (\mathsf{A} \cap X)|$. Again we begin with $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) = \frac{|X \cap (\mathsf{O} \cup \mathsf{A})|}{\max(|X|, |\mathsf{O} \cup \mathsf{A}|)} = \frac{n+k}{s+l+k}$ and slightly modified (due to $r > s$), $sim_{FS}(X, \mathsf{O}) = \frac{|X \cap \mathsf{O}|}{\max(|X|, |\mathsf{O}|)} = \frac{n}{r}$, so $sim_{FS}(X, \mathsf{O} \cup \mathsf{A}) \geq sim_{FS}(X, \mathsf{O})$ means $\frac{n+k}{s+l+k} \geq \frac{n}{r}$.

Now we use the condition $|\mathsf{O} \cup (\mathsf{A} \cap X)| > |X| > |\mathsf{O}|$ which implies $s < r < s+k < s+k+l$, i.e. $s+k-r > 0$. Simple algebra yields $\frac{n+k}{s+l+k} \geq \frac{n}{r} \iff kr \geq nl + n(s+k-r)$.

Similarly to the first case, ks is greater than the sum, and both terms are positive, so ks must be greater than nl alone. $nl + n(s + k - r) \iff ks > nl \iff \frac{k}{l} > \frac{n}{r} = sim_{FS}(X, \mathsf{O})$.

(2) First define $|X| = r$, $|\mathsf{O}| = s$, $|X \cap \mathsf{O}| = n$, $|X \cup \mathsf{O}| = m$, $|\mathsf{B} \setminus X| = l$ and $|\mathsf{B} \cap X| = k$. Clearly r, s are bigger than zero, and additionally, by Corollary 2(3), n, m, l, k are all bigger than zero.

Also note that $|B| = l + k$. Three cases emerge depending on the relative sizes of X, O, and $X \setminus \mathsf{O}$ because of the maximum function.

Case 1: $|X| \leq |\mathsf{O} \setminus \mathsf{B}| < |\mathsf{O}|$. We begin with $sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) = \frac{|X \cap (\mathsf{O} \setminus \mathsf{B})|}{\max(|X|, |\mathsf{O} \setminus \mathsf{B}|)} = \frac{n-k}{s-l-k}$ and $sim_{FS}(X, \mathsf{O}) = \frac{|X \cap \mathsf{O}|}{\max(|X|, |\mathsf{O}|)} = \frac{n}{s}$, so $sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) \leq sim_{FS}(X, \mathsf{O})$ means $\frac{n-k}{s-l-k} \leq \frac{n}{s}$. Cross multiplying and canceling terms yields $nl + nk \leq sk$. Since both terms are positive, we have $nl \leq sk \iff \frac{k}{l} \geq \frac{n}{s} = sim_{FS}(X, \mathsf{O})$ so this case is proved.

Case 2: $|\mathsf{O} \setminus \mathsf{B}| < |X| < |\mathsf{O}|$. Start with $sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) = \frac{|X \cap (\mathsf{O} \setminus \mathsf{B})|}{\max(|X|, |\mathsf{O} \setminus \mathsf{B}|)} = \frac{n-k}{s-l-k} = \frac{n-k}{s+r-s} = \frac{n-k}{r}$ and $sim_{FS}(X, \mathsf{O}) = \frac{|X \cap \mathsf{O}|}{\max(|X|, |\mathsf{O}|)} = \frac{n}{s}$. In this case though, we can only reduce our equations to the following equivalent formulations: $2sn < n^2 + mn + sk \equiv sn < sk + rn$ and then no further reductions can be done. It seems however, that we could eliminate the requirement that $\mathsf{B} \in \mathbb{B}$, so that we can restrict B to a singleton component set. Then since $|\mathsf{O} \setminus \mathsf{B}|$ could only be $|\mathsf{O}| - 1$, we could use the first case instead.

Case 3: $|\mathsf{O} \setminus \mathsf{B}| < |\mathsf{O}| < |X|$. Again we start with $sim_{FS}(X, \mathsf{O} \setminus \mathsf{B}) = \frac{|X \cap (\mathsf{O} \setminus \mathsf{B})|}{\max(|X|, |\mathsf{O} \setminus \mathsf{B}|)} = \frac{n-k}{s-l-k} = \frac{n-k}{r}$ and now $sim_{FS}(X, \mathsf{O}) = \frac{|X \cap \mathsf{O}|}{\max(|X|, |\mathsf{O}|)} = \frac{n}{r}$, and actually since k is positive, this becomes trivial as it reduces to $\frac{n-k}{r} < \frac{n}{r}$, so this case is proved.

□

Example 5 illustrates the case $|\mathsf{O} \cup (\mathsf{A} \cap X)| < |X| < |\mathsf{O} \cup \mathsf{A}|$, while Example 6 illustrates the case $|\mathsf{O} \cup \mathsf{A}| < |X|$. However the 'Majority Rule', i.e. an equivalence of Corollary 7, is valid for the Fuzzy Set index.

Corollary 8. *Let* $X \subseteq U$, $\mathsf{O} \in \mathbf{IA}(X)$, $\mathbf{x} \in \mathfrak{B}(X)$, *and* $\mathbf{x} \cap \mathsf{O} = \emptyset$. *Then:*

$$|\mathbf{x} \cap X| \geq |\mathbf{x} \setminus X| \iff \tfrac{|\mathbf{x} \cap X|}{|\mathbf{x}|} \geq \tfrac{1}{2} \implies sim_{FS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{FS}(X, \mathsf{O}).$$

Proof. If $|X| < |\mathsf{O}|$ or $|\mathsf{O} \cup (\mathsf{A} \cap X)| > |X| > |\mathsf{O}|$, then it is a direct consequence of Lemma 3. If $|\mathsf{O} \cup \mathsf{A}| < |X|$, then $sim_{FS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{FS}(X, \mathsf{O})$, regardless of the value of the ratio $\frac{|\mathbf{x} \cap X|}{|\mathbf{x}|}$. The only remaining case is $|\mathsf{O} \cup (\mathsf{A} \cap X)| < |X| < |\mathsf{O} \cup \mathsf{A}|$. Assume again that $|X| = r$, $|\mathsf{O}| = s$, $|X \cap \mathsf{O}| = n$, $|X \cup \mathsf{O}| = m$, $|\mathsf{A} \setminus X| = l$ and $|\mathsf{A} \cap X| = k$. Clearly r, s are bigger than zero, and additionally, by Corollary 2(3), n, m, l, k are all bigger than zero. Also note that $|A| = l + k$, and, since $\mathsf{A} \cap \mathsf{O} = \emptyset$, $|\mathsf{O} \cup \mathsf{A}| = |\mathsf{O}| + |\mathsf{A}| = s + l + k$ and $|X \cap (\mathsf{O} \cup \mathsf{A})| = |X \cap \mathsf{O}| + |X \cap \mathsf{A}| = n + k$. Furthermore $r > n$. If $|\mathsf{O} \cup (\mathsf{A} \cap X)| < |X| < |\mathsf{O} \cup \mathsf{A}|$, then $s + k < r < s + k + l$. We also assume that $k = |\mathsf{A} \cap X| > l = |\mathsf{A} \setminus X|$. Hence $nr > n(s + k) = ns + nk$ and $kr > ln$, which implies $nr + kr > ns + nk + nl$. But $nr + kr > ns + nk + nl \iff \frac{n+k}{s+k+l} > \frac{n}{r} \iff sim_{FS}(X, \mathsf{O} \cup \mathbf{x}) \geq sim_{FS}(X, \mathsf{O})$. □

5.6 Examples

In this section we provide several more examples which illustrate how our algorithm functions.

House	*Price ($)*	*Equiv. class*	*Qual. Index μ*
h_1	289,000	e_1	*502*
h_2	389,000	e_5	*869*
h_3	319,000	e_2	*611*
h_4	333,000	e_3	*723*
h_5	388,000	e_5	*937*
h_6	284,000	e_1	*399*
h_7	339,000	e_3	*585*
h_8	336,000	e_3	*650*
h_9	345,000	e_4	*834*
h_{10}	311,000	e_2	*366*
h_{11}	319,000	e_2	*512*
h_{12}	312,000	e_2	*622*

Class	*Elements*	*Range ($)*
e_1	h_1, h_6	*280-299,999*
e_2	$h_3, h_{10}, h_{11}, h_{12}$	*300-319,999*
e_3	h_4, h_7, h_8	*320-339,999*
e_4	h_9	*340-359,999*
		360-379,999
e_5	h_2, h_5	*380-400,000*

Table 5.1: Pawlak's space of houses and their prices. The column 'Quality Index' is used only in Example 8.

Our first example will use the Jaccard index, i.e. a special case of Marczewski-Steinhaus index with $\mu(X) = |X|$.

Example 7. *We define our universe of elements labeled $U = \{h_1, \ldots, h_{12}\}$ to be an assortment of houses, each with a price or value associated with it, as shown in Table 5.1. Based on its price, each house belongs to a representative equivalence class as demonstrated in the second table. Our classes will be defined by each range of $\$20,000$, starting from $\$280,000$ and ending with $\$400,000$ (empty classes are excluded because $\emptyset \notin \mathfrak{Comp}$). We could say that all of the houses in each class are roughly equivalent in price.*

Suppose we wish to select a subset which we are interested in. If houses $\mathsf{H} = \{h_1, h_3, h_8, h_9\}$ meet our requirements we could say that we have the financing available for each of the equivalence classes that those houses belong to. The upper and lower approximations are easy to see: $\underline{\mathbf{A}}(\mathsf{H}) = e_4$ and $\overline{\mathbf{A}}(\mathsf{H}) = e_1 \cup e_2 \cup e_3 \cup e_4$. Additionally,

the border set is $\mathfrak{B}(\mathsf{H}) = \text{comp}(\overline{\mathbf{A}}(\mathsf{H})) \setminus \text{comp}(\underline{\mathbf{A}}(\mathsf{H})) = \{e_1, e_2, e_3\}$*, and the set of intermediate approximations is* $\mathbf{IA}(\mathsf{H}) = \{\underline{\mathbf{A}}(\mathsf{H}), A_1, A_2, A_3, A_4, A_5, A_6, \overline{\mathbf{A}}(\mathsf{H})\}$ *where* $A_1 = e_1 \cup e_4$*,* $A_2 = e_2 \cup e_4$*,* $A_3 = e_3 \cup e_4$*,* $A_4 = e_1 \cup e_2 \cup e_4$*,* $A_5 = e_1 \cup e_3 \cup e_4$*, and* $A_6 = e_2 \cup e_3 \cup e_4$*. The similarity between each of these sets and* H *is as follows:* $sim_J(\mathsf{H}, \underline{\mathbf{A}}(\mathsf{H})) = \frac{|\mathsf{H} \cap \underline{\mathbf{A}}(\mathsf{H})|}{|\mathsf{H} \cup \underline{\mathbf{A}}(\mathsf{H})|} = \frac{1}{4}$*,* $sim_J(\mathsf{H}, \overline{\mathbf{A}}(\mathsf{H})) = \frac{|\mathsf{H} \cap \overline{\mathbf{A}}(\mathsf{H})|}{|\mathsf{H} \cup \overline{\mathbf{A}}(\mathsf{H})|} = \frac{2}{5}$*, and* $sim_J(\mathsf{H}, A_1) = \frac{2}{5}$*,* $sim_J(\mathsf{H}, A_2) = \frac{2}{7}$*,* $sim_J(\mathsf{H}, A_3) = \frac{1}{3}$*,* $sim_J(\mathsf{H}, A_4) = \frac{3}{8}$*,* $sim_J(\mathsf{H}, A_5) = \frac{3}{7}$*,* $sim_J(\mathsf{H}, A_6) = \frac{2}{7}$*. From all these Jaccard index values,* $\frac{3}{7}$ *is the biggest number, so* $\mathbf{Opt}(\mathsf{H}) = \{A_5\} = \{e_1 \cup e_3 \cup e_4\}$*.*

What about our algorithm? We have $\mathfrak{B}(\mathsf{H}) = \{e_1, e_2, e_3\}$*, and* $\rho(e_1, \mathsf{H}) = 1$*,* $\rho(e_2, \mathsf{H}) = \frac{1}{3}$*, and* $\rho(e_3, \mathsf{H}) = \frac{1}{2}$*. Hence* $\rho(e_1, \mathsf{H}) > \rho(e_3, \mathsf{H}) > \rho(e_2, \mathsf{H})$*, so we rename the elements of* $\mathfrak{B}(\mathsf{H})$ *as* $e_1 = \mathbf{x}_1$*,* $e_3 = \mathbf{x}_2$*,* $e_2 = \mathbf{x}_3$*. Clearly* $\rho(\mathbf{x}_1, \mathsf{H}) = 1 > sim_J(\mathsf{H}, \underline{\mathbf{A}}(\mathsf{H})) = \frac{1}{4}$ *and* $\rho(\mathbf{x}_3, \mathsf{H}) = \frac{1}{3} < sim_J(\mathsf{H}, \overline{\mathbf{A}}(\mathsf{H})) = \frac{2}{5}$*, so neither step (4) nor (5) hold, so we go to the step (6), which is the most involved.*

We begin by setting $\mathsf{O}_0 = \underline{\mathbf{A}}(\mathsf{H}) = e_4$*. Since* $sim_J(\mathsf{H}, \mathsf{O}_0) = \frac{1}{4} < sim_J(\mathsf{H}, \mathsf{O}_0 \cup \mathbf{x}_1) = \frac{2}{5}$*, we have* $\mathsf{O}_1 = \mathsf{O}_0 \cup \mathbf{x}_1 = e_1 \cup e_4$*, and since* $sim_J(\mathsf{H}, \mathsf{O}_1) = \frac{2}{5} < sim_J(\mathsf{H}, \mathsf{O}_1 \cup \mathbf{x}_2) = \frac{3}{7}$*, we have* $\mathsf{O}_2 = \mathsf{O}_1 \cup \mathbf{x}_2 = e_1 \cup e_3 \cup e_4$*. However* $sim_J(\mathsf{H}, \mathsf{O}_2) = \frac{3}{7} < \rho(\mathbf{x}_2, \mathsf{H}) = \frac{1}{2}$*, so* $\mathsf{O}_1 \notin \mathbf{Opt}(\mathsf{H})$*. Since* $sim_J(\mathsf{H}, \mathsf{O}_2) = \frac{3}{7} > sim_J(\mathsf{H}, \mathsf{O}_2 \cup \mathbf{x}_3) = \frac{2}{5}$*, we set* $\mathsf{O}_3 = \mathsf{O}_2$*. Now we have* $sim_J(\mathsf{H}, \mathsf{O}_3) = sim_J(\mathsf{H}, \mathsf{O}_2) = \frac{3}{7} > \rho(\mathbf{x}_3, \mathsf{H}) = \frac{1}{3}$*, which means that* $\mathsf{O}_2 = \{h_1, h_4, h_6, h_7, h_8, h_9\} \in \mathbf{Opt}(\mathsf{H})$*.*

Note also that $\mathsf{O}_1 = A_1$*, and* $\mathsf{O}_2 = A_5$*, and* $\mathbf{Opt}(\mathsf{H}) = \{O_2\}$*.* □

The second example uses the Marczewski-Steinhaus μ-index where μ is not cardinality.

Example 8. *Consider the same universe* $U = \{h_1, \ldots, h_{12}\}$*, the same equivalence*

classes $\{e_1, e_2, e_3, e_4\}$*, and the same set* $\mathsf{H} = \{h_1, h_3, h_8, h_9\}$ *as in the previous example. Realizing that price is only one of the factors (even though often the most important), a real estate agency, 'Best Choice,' introduced a service for customers where they will determine a quality index* μ *ranging from 0 to 100, which takes into account price, age, type of house, style, appearance, and special customer preferences. Suppose that the index values for a particular customer are described in the right column of the left part of Table 5.1. The index* μ *is extended to sets of houses* X *so we can use it to calculate the intermediate similarity values. It is defined as* $\mu(X) = \sum_{h \in X} \mu(h)$*. Clearly the index* μ *is an element-wise null-free measure as discussed in Chapter 3, so it can be used in formulas describing similarity indexes.*

What is an optimal approximation of H *with Marczewski-Steinhause index* $sim_{MS}(X, Y) = \frac{\mu(X \cap Y)}{\mu(X \cup Y)}$*? To measure the similarity between* H *and its lower approximation, we have* $sim_{MS}(\mathsf{H}, \underline{\mathbf{A}}(\mathsf{H})) = \frac{\mu(\mathsf{H} \cap \underline{\mathbf{A}}(\mathsf{H}))}{\mu(\mathsf{H} \cup \underline{\mathbf{A}}(\mathsf{H}))} = \frac{\mu(h_9)}{\mu(\{h_1, h_3, h_8, h_9\})} = \frac{\mu(h_9)}{\mu(h_1)+\mu(h_3)+\mu(h_8)+\mu(h_9)} = \frac{834}{2897} = 0.28788$*. The rest of the similarity values calculated in the same manner are as follows:* $sim(\mathsf{H}, A_1) = 0.49636$*,* $sim(\mathsf{H}, A_2) = 0.32863$*,* $sim(\mathsf{H}, A_3) = 0.33689$*,* $sim(\mathsf{H}, A_4) = 0.468515$*,* $sim(\mathsf{H}, A_5) = 0.47585$*,* $sim(\mathsf{H}, A_6) = 0.35478$*, and* $sim(\mathsf{H}, \overline{\mathbf{A}}(\mathsf{H})) = 0.4595$*. By inspection, we see the largest value is a result of comparing* H *to* $A_1 = e_1 \cup e_4$*, which is clearly different from our previous example where* A_5 *returned the largest value.*

Returning to the proposed algorithm, we have $\mathfrak{B}(\mathsf{H}) = \{e_1, e_2, e_3\}$*, and now with a different measure* μ *we calculate* ρ *for each element* e *in the border as*

$$\rho(e, H) = \frac{\mu(e \cap \mathsf{H})}{\mu(e \setminus \mathsf{H})} = \frac{\Sigma_{h \in e \cap \mathsf{H}} \mu(h)}{\Sigma_{h \in e \setminus \mathsf{H}} \mu(h)}.$$

We get $\rho(e_1, \mathsf{H}) = \frac{\mu(h_1)}{\mu(h_6)} = 2.0100$*,* $\rho(e_2, \mathsf{H}) = \frac{\mu(h_3)}{\mu(h_{10})+\mu(h_{11})+\mu(h_{12})} = 0.4073$*, and*

$\rho(e_3, \mathsf{H}) = \frac{\mu(h_4)}{\mu(h_7)+\mu(h_8)} = 0.5038$. *Hence,* $\rho(e_1, \mathsf{H}) > \rho(e_3, \mathsf{H}) > \rho(e_2, \mathsf{H})$, *as in the previous example so we again rename the elements of* $\mathfrak{B}(\mathsf{H})$ *as* $e_1 = \mathbf{x}_1$, $e_3 = \mathbf{x}_2$, $e_2 = \mathbf{x}_3$. *Since* $\rho(\mathbf{x}_1, \mathsf{H}) > sim_{MS}(\mathsf{H}, \underline{\mathbf{A}}(H))$ *and* $\rho(\mathbf{x}_3, \mathsf{H}) < sim_{MS}(\mathsf{H}, \overline{\mathbf{A}}(H))$, *steps (4) and (5) are not satisfied, so we move to step (6).*

We begin with $\mathsf{O}_0 = \underline{\mathbf{A}}(\mathsf{H}) = e_4$, *and* $\mathsf{O}_1 = \mathsf{O}_0 \cup \mathbf{x}_1 = e_1 \cup e_4$. *Note* $\rho(\mathbf{x}_1, \mathsf{H}) = 2.0100 > sim_{MS}(\mathsf{H}, \mathsf{O}_1) = 0.49636$. *So we stop here.*

If we continued, we would examine $\mathsf{O}_2 = \mathsf{O}_1 \cup \mathbf{x}_2 = e_1 \cup e_3 \cup e_4$ *and find* $sim_{MS}(\mathsf{H}, \mathsf{O}_2) = 0.47585$ *and* $\rho(\mathbf{x}_2, \mathsf{H}) = 0.5038$, *so the outcome would be the same.*

Hence for this measure μ, $\mathbf{Opt}(\mathsf{H}) = \{O_1\}$.

So it is apparent that we can use any method to evaluate size provided the range of values returned can be mapped to the natural numbers.

□

Chapter 6

Conclusion and Future Work

In this thesis, we have reviewed literature regarding rough set theory, studies in similarity, and measure theory. A definition of an *Optimal Appoximation* was provided, and generalizations of current popular similarity theories were explored. These theories were then used to craft an algorithm to determine the optimal rough set approximation(s) to a given non-exact or non-definable set. To prove this algorithm works, axioms of similarity were defined, several similarity indexes were used and investigated, and then which indexes are consistent was shown. Then the indexes were generalized to be applicable using any measurement of size instead of only number of elements. This generalization required a specification of Measure Theory where everything that is measured to have size zero is the empty set. The border and border sets were defined, as well as similarity ratio ρ, all of which led to Lemma 1. Then, Lemma 1 was used as a crucial part of the proof of correctness of Theorem 1 which validated the usefulness of the function ρ in determining an optimal approximation. Next, the vital results of Theorem 1 were used (as well as Lemma 1, and an

algebraically trivial Lemma 2) to prove that Theorem 2 does indeed find the greatest optimal approximations to a target set. If equal optimal approximations exist, the theorem can show these by retracing each previous intermediate approximation until the similarity decreases. Lastly, the Theorem 2 was transformed into a set of instructions forming Algorithm 1.

While studying the properties of similarity measures and optimal approximations, several possible open avenues of research arose. One topic to investigate is whether any version of the Triangle Inequality applies. If $sim(A, B) < sim(B, C)$ and $sim(B, C) < sim(C, D)$ then there might be some relationship between $sim(A, B)$ and $sim(C, D)$ beyond the obvious transitivity fact that $sim(A, B) < sim(C, D)$ or possibly some relation between similarity of the other sets $sim(A, C)$, $sim(B, D)$, and $sim(A, D)$. Alternatively, the fact that $sim(A, B) < sim(B, C)$, might imply some value for $sim(A, C)$.

As mentioned in [35], to extend any similarity measure, the accuracy could be increased by adding more functions and possibly more variables to capture more information about the subject, referent, universe, and type of comparison being made. This is certainly an area rich in properties to investigate.

There is also Pawlak's notion of rough inclusion that could be investigated with respect to optimal approximations. If one set is roughly included in another, is there a relationship between the optimal approximations of each set? Or a relationship between an optimal approximation and a roughly included set?

The idea of special inclusion operators leads to the idea of an optimal inclusion operator, which could represent if a set is a part of all optimal approximations which involve it, or if a set is a part of the optimal approximation of another set. Knowledge

of the results of these operators could be used to accelerate our algorithm, but this exploration is left for future study.

The assumption of similarity should also be examined. There are both instances in which our assumption of symmetry is not valid, and a variety of applications which assume symmetry, such as any example where elements of the universe can be regarded as euclidean points. If only numerical data are present, a euclidean foundation is natural, though dimensions can still be regarded as attributes, and arbitrary equivalence classes created from them.

In another vein, the concept of measures was used to generalize the evaluation of the size of a set. Another generalization could be used to refer to the granularity of which we are focused on. Using the example of similarity of organisms in their environments, we could find that two types of animal are very similar at a high level due to having similar prey, eating habits, activity, routines, but if we look much closer, we might find large differences between the species, and each is more similar to another type based on different criteria. This could alternatively be done by using a different equivalence relation on the universe to induce difference classes.

Another idea that should be checked, is if the algorithm is still applicable when using measures to evaluate the size of each element of a set individually instead of using the size of the whole set. It is expected to remain useful since additivity is assumed in the geometric setting. Alternatively, measuring the similarity between all elements pairwise, rather than evaluating the size of an entire set or similarity between sets could lead to further insight.

In conclusion, the pursuit of new approximation algorithms, or refinements to this one seems like a vast expanse of open problems to investigate. Topics such as these

could yield some very useful data organization or filtration techniques.

Bibliography

[1] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 3rd edition, 1995. ISBN 0-471-00710-2.

[2] Josias Braun-Blanquet. *Pflanzensoziologie*. Springer, 1928.

[3] Victor Bryant. *Metric spaces: iteration and application*. Cambridge University Press, 1985.

[4] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2012.

[5] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[6] Ivo Düntsch. Rough sets and algebras of relations. *Incomplete Information: Rough Set Analysis*, 13:95, 2013.

[7] Maurice René Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo(1884-1940)*, 22(1):1–72, 1906.

[8] Wendell Garner and Donna Sutliff. The effect of goodness onencoding time in visual pattern discrimination. *Perception & Psychophysics*, 16(3):426–430, 1974.

[9] Paul Halmos. *Measure Theory*. Van Nostrand, 1950.

[10] Paul Halmos. Measure theory. *Graduate Texts in Mathematics*, v.18, 1974.

[11] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37: 547–549, 1901.

[12] Ryszard Janicki. On rough sets with structures and properties. *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pages 109–116, 2009.

[13] Ryszard Janicki and Adam Lenarčič. Optimal approximations with rough sets. *Rough Sets and Knowledge Technology*, pages 87–98, 2013.

[14] Ryszard Janicki and Adam Lenarčič. Optimal approximations with rough sets and similarities in measure spaces. *International Journal of Approximate Reasoning*, 71:1–14, 2016.

[15] Jon Kleinberg and Éva Tardos. *Algorithm design*. Addison-Wesley, 2006.

[16] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[17] Casimir Kuratowski. *Topology*, volume 1. Elsevier, 2014. ISBN 9781483272566. URL `https://books.google.ca/books?id=WZLOBQAAQBAJ`.

[18] Edward Marczewski and Hugo Steinhaus. On a certain distance of sets and the corresponding distance of functions. In *Colloquium Mathematicae*, volume 6, pages 319–327. Institute of Mathematics Polish Academy of Sciences, 1958.

[19] Edward Marczewski and Hugo Steinhaus. O odległości systematycznej biotopów. *Applicationes Mathematicae*, 3(4):195–203, 1959.

[20] Marshall E. Munroe. Introduction to measure and integration. *AMC*, 10:12, 1953.

[21] Zdzisław Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, 1982.

[22] Zdzisław Pawlak. Rough sets and fuzzy sets. *Fuzzy sets and Systems*, 17(1): 99–102, 1985.

[23] Zdzisław Pawlak. *Rough sets: Theoretical aspects of reasoning about data*, volume 9. Kluwer, Dordrecht, 1991.

[24] Zdzisław Pawlak. Rough set approach to knowledge-based decision support. *European journal of operational research*, 99(1):48–57, 1997.

[25] James F Peters, Andrzej Skowron, Piotr Synak, and Sheela Ramanna. Rough sets and information granulation. In *International Fuzzy Systems Association World Congress*, volume 2715, pages 370–377. Springer, 2003.

[26] Hassan Rezaei, Masashi Emoto, and Masao Mukaidono. New similarity measure between two fuzzy sets. *JACIII*, 10(6):946–953, 2006.

[27] Eleanor Rosch. Cognitive reference points. *Cognitive psychology*, 7(4):532–547, 1975.

[28] Jamil Saquer and Jitender S Deogun. Concept approximations based on rough sets and similarity measures. *Applied Mathematics and Computer Science*, 11 (3):655–674, 2001.

[29] Berthold Schweizer and Abe Sklar. Statistical metric spaces. *Pacific Journal of Mathematics*, 10(1):313–334, 1960.

[30] Michael Sipser. *Introduction to the Theory of Computation*, volume 2. Thomson Course Technology Boston, 2006.

[31] Lennart Sjöberg. A cognitive theory of similarity. *Göteborg Psychological Reports*, 2(10), 1972. URL `https://books.google.ca/books?id=mBE_NAAACAAJ`.

[32] Dominik Slezak and Wojciech Ziarko. The investigation of the bayesian rough set model. *International Journal of Approximate Reasoning*, 40(1-2):81–91, 2005.

[33] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologishe Skrifter*, 5:1–34, 1948.

[34] Amos Tversky. Features of similarity. *Psychological Reviews*, 84(4):327–352, 1977.

[35] Amos Tversky and Itamar Gati. Studies of similarity. *Cognition and categorization*, 1(1978):79–98, 1978.

[36] Amos Tversky and Itamar Gati. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

[37] Yiyu Yao. A comparative study of fuzzy sets and rough sets. *Information sciences*, 109(1-4):227–242, 1998.

[38] Yiyu Yao and Tao Wang. On rough relations: An alternative formulation. *Lecture notes in computer science*, pages 82–90, 1999.

[39] Lotfi Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.