دانشکده ریاضی و رایانه
بخش ریاضی

رساله برای دریافت درجه دکتری ریاضی

دسته بندی بر اساس تشابه و عدم تشابه

اساتید راهنما:

دکتر ماشاله ماشین چی

دکتر محمد علی دهقان

مؤلف:
محمود کشاورزی

دی ماه 1389

دانشگاه شهید باهنر کرمان

دانشکده ریاضی و رایانه
بخش ریاضی

رساله برای دریافت درجه دکتری ریاضی

دسته بندی بر اساس تشابه و عدم تشابه

اساتید راهنما:

دکتر ماشاله ماشین چی

دکتر محمد علی دهقان

مؤلف:
محمود کشاورزی

دی ماه 1389

To my very dear parents,

my dear wife,

and my dear sons.

Table of Contents

Abstract

The concepts of similarity and dissimilarity have been the interest of many researchers. Basically, in the studies the similarity between two objects or phenomena, has been discussed. In this thesis, we consider the case when the resemblance or similarity among three objects or phenomena of a set, 3-similarity in our terminology, is desired. Later we will extend our definitions and propositions to n-similarity. Since in some cases recognizing dissimilarity is easier than similarity, we try to find a connection between these relations based on specific functions. We will also define the concept of $\lambda - cuts$ and relations in connection with the concepts of 3-similarity and 3-equivalence relations. It should mentioned that the related definitions and propositions are on the basis of the genaralization of the concepts of the $\lambda - cuts$ and 3-equivalence relations. At the end, some applications as well as their algorithms will be presented to support our ideas and propositions.

Acknowledgements

I wish to express my special gratitude and thanks to Prof. Mashaallah Mashinchi and especially Prof. Mohammd Ali Dehghan for their invaluable guidance and great help for this thesis. I would like to thank Prof. Mahmoud Mohseni, Prof. Hamid Mazaheri, and Prof. Faramarz Sadeghi who read this thesis and gave useful comments.

Special thanks to my dear wife, Fereshteh, and my beloved sons, Kian and Behnam for their patience during the years of my studies.

I also would like to thank Prof. Alireza Bahrampour, Prof. Nasrolah Gerami, Prof. Mahbanou Tata, Prof. Reza Ranjbar Karimi, Prof. Hasan Ranjbar Askari, Prof. Ahmad Abdollahzadeh (of Amir Kabir University), Prof. Seyed Naser Hossaini, and Miss. Tahereh Karimi-afshar. All these people guided me or helped me in a way or the other in connection with my studies. I am also indebted to the department of mathematics of Shahid Bahonar University of Kerman and Vali Asr University of Rafsanjan for their support towrds getting my Ph.D. degree.

Kerman, Iran
December 27, 2010

Mahmood Keshavarzi

Introduction

0.1 Background

In the real world we sometimes encounter situations where we have to find *similarity* between two or more objects, concepts, or phenomena and classify them into groups for some predefined reason, or to study their common properties. The degree to which these objects, concepts, or phenomena are similar or compatible (dissimilar in some cases) is a basic component of human reasoning and consequently is very important in the development of automated classification, diagnosis, and decision systems. The notion of similarity plays an important role in theories of knowledge and behaviour and has been dealt with in pattern recognition [13], decision making [64], psychology literature [28], similarity-based clustering approaches [33, 69], and approximate reasoning [8, 20, 43, 61]. Also, the performance of different similarity measures have been used to evaluate document summarizations [2]. Artificial Intelligence experts have studied the computational similarity models as a new method for information retrieval [59]. In all these research and discussions on similarity relations, similarity between two objects, or concepts have been studied. In this thesis we will look at the concept of similarity from different perspectives, and that is the similarity and resemblance among three or more objects, in our terminology, 3-similarity, 4-similarity, or in general n-similarity. That is, we propose a way to select three similar objects (or n-similar) out of a group of objects. Through some examples on real data, we will show how our propositions and ideas can be applied in the real world. Also, we will show that by using some functions we can convert similarities to *dissimilarities* and make proper decisions based on them in a more realistic manner. Then we propose the concept of the 3-similarity, which is the similarity of three objects among a set of objects. That is, we would like to find out which three members, out of a group of members, have the most resemblance and similarity to each other. Thereafter, we continue with stating the definitions, propositions, theorems, and lemmas regarding 3-similarity relations and equivalence 3-relations. We will show that all the definitions and propositions related to 2-similarity relations could also be stated for 3-similarity relations. Analogies and differences between the 2-similarity and 3-similarity will be studied. Also, we will show that if we have a 2-similarity, under certain conditions, 3-similarity could be obtained. Finally we will show that the idea can be generalized towards the n-similarity relations. In the third chapter we will review the classification based on 3-similarity and n-similarity that had been discussed in [37]. We will extend the definitions and propositions to the concept of 3-dissimilarity, and in general n-dissimilarity relations. Also, we will extend the relationship between similarity and dissimilarity to 3-similarity and 3-dissimilarity relations in chapter 4. Under the section of "Applications, Algorithms, and Working flows", i.e. the fifth chapter, we have developed some algorithms to be used on real data to support our propositions and ideas. In fact there are three applications that have been discussed, the first one shows, through an example, how we can select, out of a set of objects, three of them that are most similar, provided their pairwise similarities are known before hand.

0.2 Organization of this thesis

Chapter 1, states a brief history of similarity and similarity measures and different approaches concerning them. In chapter 2 a comprehensive definition of equivalence relations and equivalence classes has been brought. Chapter 3 discusses the concept of classification based on similarities. Chapter 4 discusses classification based on dissimilarity measures. Chapter 5 studies equivalence relations on similarities and dissimilarity, as well as n-similarities and n-dissimilarity. In chapter 6 we will study the relationship between similarities and dissimilarity and proposes a formula by which we can convert similarity and dissimilarity into each other. In chapter 7 through some examples we will show how our ideas and propositions can be applied in the real world. Chapter 8 is the conclusions and future works.

Chapter 1

The Concept of Similarity

1.1 Introduction

Our aim, by studying the materials in this chapter is to looking at the concept of similarity from different perspectives and angles. No need to mention that nowadays many researchers are working on the concept of similarity [10, 70, 67, 31, 49, 50]. Any measurement of similarity of objects will be based on certain assumptions concerning the properties of their relation. Sometimes the degree of similarity between objects should be determined relative to a given context or procedure. Here we begin with studying the different ways to measure similarity. Then some approaches to similarity will be looked at. At the end, fuzzy sets and similarity will be discussed since approximate and imprecise information are represented by introducing a similarity relation which is a mathematical tool that allows to weak the crisp notion of equality.

1.2 Similarity and Similarity Measures

There are different interpretation of similarity in different scientific disciplines. For instance in *Geometry*, two objects are similar if they both have the same shape. In *Chemistry*, chemical similarity (or molecular similarity) refers to the similarity of chemical elements, molecules or chemical compounds with respect to either structural or functional qualities, i.e. the effect that the chemical compound has on reaction partners in an organic or biological settings [35, 46]. *Semantic Similarity*[1, 34] is a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content. The main challenge in semantic similarity measurement is the comparison of meanings. In essence, semantic similarity, semantic distance, and semantic relatedness all mean, "How much does term A have to do with term B?" The answer to this question, as given by the many automatic measures of semantic similarity/relatedness, is usually a number, between -1 and 1, or between 0 and 1, where 1 signifies extremely high similarity/relatedness, and 0 signifies little-to-none. Since the theme of the applications and examples of this thesis is more related to psychological approaches to similarity, we would like to enter into this discussion in more details as in the following section.

1.3 Psychological approaches to similarity

Human judgments of similarity have been subject to research in psychology for more than fifty years [25]. In fact the notion of similarity originated in psychology and was established to determine why and how entities are grouped to categories, and why some categories are comparable to each other while others are not [25, 26]. In social psychology, similarity refers to how closely interests, attitudes, values, and personality match between people. Research has consistently shown that similarity leads to interpersonal attraction. Many forms of similarity have been shown to increase liking. Similarities in opinions, interpersonal styles, amount of communication skill, demographics, and values have all been shown in experiments to increase liking.

Several explanations have been offered to explain similarity increases interpersonal attraction. People with similar interests tend to put themselves into similar types of settings. For example, two people interested in literature are likely to run into each other in the library and form a relationship. A very important point to notice here is the way similarity should be measured in psychological approaches. There are two approaches that we will consider, Featural and structural. We briefly take look at the both approaches here.

1.3.1 Featural Approaches

Featural approaches [66] were developed in order to cope with the limitations of the mental distance approaches. For example, spaces are symmetric. The distance between two points is the same, no matter which point you start from. However, psychological similarity is not symmetric. For instance, saying "That surgeon is a butcher" means something quite different from saying "That butcher is a surgeon". Tversky in [66] noted the assessment of similarity may be better described by comparing features than computing metric distance between points. He describes similarity as a feature matching process, similarity among objects is expressed as a linear combination of the measure of their common and distinct features [77].

In featural approaches concepts are represented by lists of features that describe properties of the items. Similarity between items then, would be comparing their feature lists. Identical features make the common points of the pair, and features that are contained in one feature set but not the other are differences of the pair. It is possible to judge that similarity between the items increases with the number of common features (weighted by the importance the common features) and decreases with the number of differences (weighted by the importance of the different features).

1.3.2 Structural Approaches

Structural approaches to similarity [22] were emerged to compensate the limitations of the featural approaches. In particular, in featural approaches we assumed that the common and uncommon features of items are independent of each other. Whereas, they are not psychologically independent. In fact, determining the differences between two items requires finding the commonalities. For example, if we compare a car and a motorcycle. Both have wheels. That is a commonality. However, cars have four wheels, while motorcycles have two wheels. That is a difference. Because this difference required first finding a commonality between the pair, it is called an alignable difference. Alignable differences contrast with nonalignable differences which are aspects of one concept that have no correspondence in the other. For example, cars have seat belts and motorcycles do not. Research suggests that alignable differences have a larger impact on people's judgments of similarity than do nonalignable differences. Thus, the relationship between the commonalities of a pair and the differences is important for understanding people's assessments of similarity. Structural approaches to similarity emerged from research on analogy.

Since measuring the exact similarity among objects is not all the time possible, we measure approximate similarities. The tool for this purpose is fuzzy mathematics. So, in the next section we take a look at relationship between fuzzy sets and similarity measures, and a brief review of the works that have been done on this topic.

1.4 Fuzzy sets and Similarity

The theory of fuzzy sets proposed by Zadeh in 1965 [74] has achieved a great success in various fields. Fuzzy set is completely non-statistical in nature, and 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. In fact the idea of describing all shades of reality was for long the obsession of some logicians [42, 60]. During last four decades the fuzzy set theory has rapidly developed into an area which scientifically as well as from the application point of view, is recognized as a very valuable contribution to the existing knowledge (see [5, 6, 15, 18, 27, 29, 51, 76]. After the emergence of fuzzy set theory, the simple task of looking at relations as fuzzy sets on the universe U was accomplished in a celebrated paper by Zadeh [75], he introduced the concept of fuzzy relation, defined the notion of equivalence. Compared with crisp relations, they have greater expressive power. They are considered as softer models for expressing the strength of links between elements.

In intelligent activities, it is often needed to compare and couple between two fuzzy concepts. That is, we need to check whether two knowledge patterns are exactly (or approximately) identical, similarity measure is a proper tool for this purpose. That is, this tool helps us to determine the grade of similarity between two groups or two elements. The measure of similarity of fuzzy sets has been proposed by Zwick et al. [77]. Some of the applications of similarity measures to fuzzy sets include, applications in fuzzy mathematics [72], decision making, market prediction, and pattern recognition [77, 11, 54]. Many other authors and researchers have proposed similarity measures of fuzzy sets [17, 30, 54, 12, 24, 71, 73] that can be viewed as generalizations of the classical set-theoretic similarity measures. Rezaei et al. in [57, 58] proposed a new similarity measure between fuzzy sets and extended it to define two other similarity measures. Similarity/dissimilarity topics have also been extensively used in the study of Intuitionistic Fuzzy Sets (IFS) [47, 16]. IFS was introduced by Atanassov [3, 4]. Rezaei and Mukaidono [55, 56] proposed new similarity measures for IFS. They also studied the relationship between similarity and dissimilarity measures of two IFS sets.

Chapter 2

Equivalence Relations and Equivalence Classes

2.1 Introduction

Since in our studies in this thesis equivalence relations and equivalence classes have an important role, we bring some of their related definitions here. Equivalence relation is defined in set theory as an important notion of mathematics, which is a mathematical concept on a given set that provides a way for elements of that set to be identified with other elements of the set, that is, considered equivalent to, for some purpose. The power of an equivalence relation lies in its ability to partition a set into the disjoint union of subsets called equivalence classes. Because of its power to partition a set, an equivalence relation is one of the most used and pervasive tools in mathematics. As far as similarity and dissimilarity relations are concerned, since similar items of some sort can reside in equivalence classes, understanding the equivalent relation and equivalent classes is important.

2.2 Equivalence Relation and Equivalence Classes

Before discussing equivalence relation, it is necessary to review the general concept of relation on a given set [19].

2.2.1 Relation

Let U be the given set. A relation, $\sim$, on U is a subset of the Cartesian product of $U \times U$. Hence any particular element, $x \in U$ has the relation $\sim$ with any element, $y \in U$ if and only if $(x, y) \in \sim$. Note $\sim$ is just a subset of the Cartesian product and the interpretation of $\sim$ is key point here.

2.2.2 Equivalence Relation

Equivalence relations are ways to partition a set into subsets of equivalent elements. Being equivalent is then interpreted as being the same, such as different views of the same object or different ordering of the same elements, etc. Equivalence relation is a special case of relation. To pave the way for our discussions on equivalence relation, we first bring definitions on equivalence relations and equivalence classes.

Definition 2.2.1. Let U be a set and x, y, and z be elements of U. An equivalence relation, $\sim$, on U is a relation on U which has the following properties:

- Reflexive: (x, x) is in $\sim$ for all $x \in U$.
- Symmetric: if (x, y) is in $\sim$, then (y, x) is in $\sim$.

- Transitive: if (x, y) and (y, z) are in $\sim$, then (x, z) is in $\sim$. Note that U might be empty, in that case, $\sim$ is empty too. If U is non-empty, then $\sim$ is non-empty as well.

The relation of equality and clockwork arithmetic are obvious examples of equivalence relations on a set like U defined above.

2.2.3 Equivalence Classes

Definition 2.2.2. Let U be a set, $\sim$ be an equivalence relation on U, and $x \in U$. The equivalence class of x is the subset of U that contains all elements of U that are equivalent to x under $\sim$. In symbols, the equivalence class of x is the subset $\{y$: $x \sim y$ and y is an element of $U\}$ of U.

The equivalence class representative is, in general, not unique. That is, if both x and y are in an equivalence class, then either one could represent that equivalence class. Every equivalence relation produces equivalence classes. For instance in the "clockwork arithmetic" relation, there are twelve equivalence classes as follows:

$$\{0, 12, ...\}, \{1, 13, ...\}, ..., \{11, 23, ...\}$$

The first equivalence class listed, $\{0, 12, ...\}$, can be called the equivalence class of 0, 12, or 24. Truly it can be called the equivalence class of any multiple of 12 including 0.

As an another example: modular arithmetic. We say an integer a is congruent to another integer b modulo a positive integer n, denoted as $a \equiv b \mod n$, if $b - a$

is an integer multiple of n. To illustrate this definition, let $n = 3$ and let U be the set of integers from 0 to 11. Then, $x \equiv y \mod 3$, if x and y both belong to $U_0 = \{0, 3, 6, 9\}$ or both belong to $U_1 = \{1, 4, 7, 10\}$ or both belong to $U_2 = \{2, 5, 8, 11\}$. This can be easily verified by testing each pair. Congruence modulo 3 is in fact an equivalence relation on U. To see this, we show that congruence modulo 3 satisfies the three required properties. *reflexive*: Since $x - x = 0$, we know that $x \equiv x \mod 3$. *symmetric*: If $x \equiv y \mod 3$ then x and y belong to the same subset U_i. Hence, $y \equiv x \mod 3$. *transitive*: Let $x \equiv y \mod 3$ and $y \equiv z \mod 3$. Hence x and y belong to the same subset U_i and so do y and z. It follows that x and z belong to the same subset. More generally, congruence modulo n is an equivalence relation on the integers.

As mentioned above, the representative of an equivalence class can be any element of that class. Can confusion arise if two equivalence classes share a common element? No. The reason is that any two different equivalence classes are disjoint. This fact can easily be gleaned in this example by looking at the equivalence classes listed above. To prove this fact in general, one needs a theorem called *partition theorem*:

Theorem 2.2.1. *[19] Let U be a non-empty set and $\sim$ an equivalence relation on U. The equivalence classes of $\sim$ form a partition (a disjoint collection of non-empty subsets whose union is the whole set) of U.*

A converse of partition theorem also exists.

Theorem 2.2.2. *[19] Let U be a set and P be a partition of U. P corresponds to an equivalence relation, $\sim$, on U where, for x and y elements of U, $x \sim y$ if and only if x and y lie in the same element of P.*