Categorization Axioms for Classification and Clustering

Jian YU
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, Beijing, China
Email: jianyu@bjtu.edu.cn

Zongben XU
School of Mathematics and Statistics
Xi’an Jiaotong University, Xi’an, China
Email: zbxs@mail.xjtu.edu.cn

Abstract

Categorization is a fundamental problem for pattern recognition, machine learning and human recognition systems, which involves classification and clustering. In this paper, we first propose a unified representation of categorization results and then present an axiomatic framework for categorization by assuming that a proper categorization result should satisfy categorization equivalency axiom, sample separation axiom and category separation axiom. The proposed axiomatic framework is not only consistent with categorization theories in cognitive science, but also classifies categorization results and partitions. The proposed separation axioms can lead to three principles of designing categorization methods. And classification and clustering are theoretically investigated based on categorization axioms.

1 Introduction

Categorization is a fundamental ability that allows people to cognize objects in the world (Lakoff, 1987). It has always been desired for a computer to possess the categorization ability of a human being. Generally, when perceiving objects in the world, a computer can categorize them into some concepts, such as tree, sky, water and cat. For a computer, categorization means that objects are grouped into categories, a process that involves classification and clustering. Classification requires that objects feature a category label; in contrast, clustering does not require objects to have a category label, except for the case of semi-supervised clustering, in which some information regarding category labels is needed (Basu et al., 2003).

In the literature, many classification methods and clustering algorithms have been developed. For classification, some theories such as PAC theory (Valiant, 1984) and statistical learning theory (Vapnik, 2000) have been proposed. For clustering, Kleinberg (2003) presented an impossibility theorem for clustering, which is considered a rigorous proof of the difficulty in finding a unified framework for different clustering approaches (Correa-Morris, 2013). To the best of our knowledge, no axiomatic framework of categorization is proposed. In this paper, a unified axiomatic framework for categorization is proposed based on categorization models in cognitive science.

The major contributions of this paper are as follows:
1) A representation for categorization results is proposed based on the definition of category (dis)similarity mapping.

2) Categorization results are axiomatized using a sample separation axiom, category separation axiom and categorization equivalency axiom based on categorization theories in cognitive science.

3) The proposed axioms can not only classify categorization results into proper categorization, overlapping categorization and improper categorization, but also classify partitions into proper partitions, overlapping partitions and improper partitions. Moreover, the boundary set is theoretically defined based on sample separation axiom.

4) The proposed axioms can follow three principles of developing categorization methods.

5) Classification and clustering are theoretically studied based on categorization axioms.

The remainder of the paper is organized as follows: In section 2, a representation of a categorization result is discussed based on categorization models in cognitive science. In section 3, sample separation axiom, category separation axiom and categorization equivalency axiom are introduced for a categorization result, and categorization results are classified based on the proposed axioms. In section 4, how to design a cate-
categorization method is discussed. In Section 5, classification and clustering are theoretically investigated respectively. The final section offers concluding remarks.

2 Representation of Categorization Results

In this paper, we will investigate how to represent a categorization result when the given data set $X = \{x_1, x_2, \ldots, x_n\}$ is categorized into $c$ categories $X_1, X_2, \ldots, X_c$.

A categorization result usually represents a set of categories. Hence, how to represent a category is the key for representing a categorization result. Transiently, a category has two kinds of representation, one derived from the intensional definition of a category, the other based on the extensional definition of a category. In this paper, the first representation is called cognitive representation, the second representation is called extensional representation.

In cognitive science, psychological research has offered four mental representations of categories (Murphy, 2004): classical theory, prototype theory, exemplar theory, knowledge theory. As for classical theory, a category $A$ can be defined by a proposition $\alpha$, which states the necessary and sufficient conditions for concept membership. As for prototype theory (Rosch, 1978), a category is represented by a prototype. As for exemplar theory (Medin and Schaffer, 1978), a category is represented by multiple exemplars. As for knowledge theory, a category is part of a general knowledge about the world (Murphy and Medin, 1985). In this paper, the mental representation of a category is called cognitive representation of a category, which is always supposed to be existed for a category even in an implicit state. For simplicity, the cognitive representation of a categorization result is represented by $X = \{X_1, X_2, \ldots, X_c\}$, where $X_i$ represents category $X_i$.

The extensional representation of a categorization result is represented by a partition matrix $U = [u_{ik}]_{c \times n}$, where $\forall i \forall k, u_{ik} \geq 0$ represents the membership of the object $x_k$ belonging to the category $X_i$. For a categorization result, its extensional representation can also be called a partitional representation.

As reported in the literature, various constraints on partition matrices result in different partitions:

**Hard Partition:** If $\sum_{i=1}^{c} u_{ik} = 1, u_{ik} \in \{0, 1\}, \forall i, 1 \leq i \leq c, \sum_{k=1}^{n} u_{ik} < n$, then $U_h = [u_{ik}]_{c \times n}$ is called hard partition.

**Soft Partition:** If $\sum_{i=1}^{c} u_{ik} = 1, 0 \leq u_{ik} \leq 1$, and $\forall i, 0 < \sum_{k=1}^{n} u_{ik} < n$. then $U_s = [u_{ik}]_{c \times n}$ is called soft partition.

**Possibilistic Partition:** If $\forall k, \sum_{i=1}^{c} u_{ik} > 0, u_{ik} \geq 0$, and $\forall i \leq i \leq c, \sum_{k=1}^{n} u_{ik} > 0$, then $U_p = [u_{ik}]_{c \times n}$ is called possibilistic partition.

Depending on the type of partition involved, a categorization method can be classified as a hard categorization method and a soft categorization method.

**Hard Categorization:** If a categorization method outputs a hard partition matrix, it is called a hard categorization method.

**Soft Categorization:** If a categorization method outputs a soft or possibilistic partition, it is called a soft categorization method.

Cognitive representation also has the capacity for categorization. In cognitive science, categorization is based on the similarity of an object to the underlying category representation. Based on this concept, a category similarity mapping can be defined by computing the similarity between objects and categories.

**Category similarity mapping:**

$Sim: X \times \{X_1, X_2, \ldots, X_c\} \rightarrow R_+$ is called category similarity mapping if an increase in $Sim(x_k, X_i)$ indicates greater similarity between $x_k$ and $X_i$ and a decrease in $Sim(x_k, X_i)$ indicates less similarity between $x_k$ and $X_i$.

Similarly, category dissimilarity mapping can be defined as follows:

**Category dissimilarity mapping:**

$Ds: X \times \{X_1, X_2, \ldots, X_c\} \rightarrow R_+$ is called category dissimilarity mapping if an increase in $Ds(x_k, X_i)$ indicates less similarity between $x_k$ and $X_i$ and a decrease in $Ds(x_k, X_i)$ indicates greater similarity between $x_k$ and $X_i$.

According to the above analysis, a categorization result can be represented by $(X, U, Sim)$ or by $(X, U, Ds)$.

3 Categorization Axioms

In common sense, a categorization result should be consistent with the human cognition. In cognitive science, one object $x$ should be assigned to a category $A$ not other categories if the similarity between $x$ and the cognitive representation of $A$ is maximal, a category should have at least one object, and the cognitive representation of a category should have the same categorization capacity as its partitional representation. Based on the above observation, it is natural to have three categorization axioms as follows:
1) Sample Separation Axiom: For any given object, there is only one category whose cognitive representation is most similar to this object.

2) Category Separation Axiom: For any given category, there is at least one object which is most similar to the cognitive representation of this category.

3) Categorization Equivalency Axiom:

For any given category, its cognitive representation should have the same categorization capacity as its partitional representation.

More accurately, if a categorization result of the data set \(X\) is represented by \((X, U, Sim)\), then two separation axioms and categorization equivalency axiom can be expressed in mathematical terms as follows.

1) Sample Separation Axiom: 
\[
\forall k \exists i \forall j((j \neq i) \rightarrow (Sim(x_k, X_i) > Sim(x_k, X_j)))
\]

2) Category Separation Axiom: 
\[
\forall i \exists k (\forall j((j \neq i) \rightarrow (Sim(x_k, X_i) > Sim(x_k, X_j))))
\]

3) Categorization Equivalency Axiom: 
\[
\forall k (\arg \max_i u_{ik} = \arg \max_i Sim(x_k, X_i))
\]

When a categorization result of the data set \(X\) is represented by \((X, U, Sim)\), categorization axioms can be expressed as follows:

1) Sample Separation Axiom: 
\[
\forall k \exists i \forall j((j \neq i) \rightarrow (Ds(x_k, X_i) < Ds(x_k, X_j)))
\]

2) Category Separation Axiom: 
\[
\forall i \exists k (\forall j((j \neq i) \rightarrow (Ds(x_k, X_i) < Ds(x_k, X_j))))
\]

3) Categorization Equivalency Axiom: 
\[
\forall k (\arg \max_i u_{ik} = \arg \min_i Ds(x_k, X_i))
\]

If separation axioms hold, some properties of a categorization result can be obtained as Theorem 1.

**Theorem 1.** If a categorization result \((X, U, Sim)\) satisfies category separation axiom, then we have two conclusions.

1. \(\forall i \exists j (i \neq j) \rightarrow (X_i \neq X_j)\).
2. There exists at least \(c\) objects \(x_k\), such that \(\forall i \exists j (i \neq j) \rightarrow (x_k \neq x_j)\).

The proposed separation axioms can classify categorization results into the three following types:

**Proper categorization result:** If a categorization result follows sample separation axiom and category separation axiom, such a categorization result is a proper categorization result.

**Overlapping categorization result:** If a categorization result obeys category separation but violates sample separation, it is called an overlapping categorization result.

**Improper categorization result:** If a categorization result violates category separation axiom, it is called an improper categorization result.

Obviously, proper categorization result and overlapping categorization result are useful in practice. However, a good categorization result is not expected to be an improper categorization result. In particular, a categorization method is not expected to generate improper categorization results when the given data set has a well categorization structure. Two special cases of improper categorization result can be defined as follows.

**Coincident categorization result:** For \(X = \{X_1, X_2, \ldots, X_c\}\), if \(\exists i \exists j (i \neq j) (X_i = X_j)\), it can be called coincident categorization result.

**Totally coincident categorization result:** For \(X = \{X_1, X_2, \ldots, X_c\}\), if \(\forall i \forall j (X_i = X_j)\), it is called totally coincident categorization result.

According to the categorization equivalency axiom, a classification of partitions can be equivalently offered according to classification of the categorization results as follows:

**Proper partition:** A partition \(U = [u_{ik}]_{c \times n}\) is proper if \(\forall k \exists i \forall j((j \neq i) \rightarrow (u_{ik} > u_{jk})\) and \(\forall i \exists k \forall j((j \neq i) \rightarrow (u_{ik} = u_{jk}))\).

**Overlapping partition:** A partition \(U = [u_{ik}]_{c \times n}\) is overlapping if \(\exists k \exists i \exists j((j \neq i) \land (u_{ik} = u_{jk}) = \max, u_{ik})\) and \(\forall i \exists k \forall j((j \neq i) \rightarrow (u_{ik} > u_{jk}))\).

**Improper partition:** A partition \(U = [u_{ik}]_{c \times n}\) is improper if \(\exists i \forall k \exists j((i \neq j) \land (u_{ik} \leq u_{jk}))\).

More detailed, several special cases of improper partitioning can be defined as follows.

**Covering partition:** If a partition \(U = [u_{ik}]_{c \times n}\) satisfies \(\exists i \exists j \forall k ((i \neq j) \rightarrow (u_{ik} \leq u_{jk}))\), \(U = [u_{ik}]_{c \times n}\) is called a covering partition.

**Coincident partition:** If a partition \(U = [u_{ik}]_{c \times n}\) satisfies \(\exists i \exists j \forall k ((i \neq j) \rightarrow (u_{ik} = u_{jk}))\), \(U = [u_{ik}]_{c \times n}\) is called a coincident partition.

**Uninformative partition:** \(U_\pi = [\pi_1, \pi_2, \ldots, \pi_c]^T \otimes 1_{1 \times n}\) is called an uninformative partition, where \(\otimes\) represents Kronecker product, \(1\) denotes the vector of all 1’s.

**Absolute uninformative partition:** \(U_{c^{-1}} = [c^{-1}]_{c \times n}\) is called an absolute uninformative partition.

When a categorization result is not proper, there are some objects theoretically belonging to two and more
categories. In other words, some objects are in the borderline of some category. Based on this fact, boundary set can be defined as follows.

**Boundary set:** If a categorization result \((X, U, Sim)\) for a data set with \(n\) objects, the boundary set for \((X, U, Sim)\) is defined as follows.

\[
BS(X, U, Sim) = \{x_k | \arg \max_{1 \leq i \leq c} Sim(x_k, X_i) > 1\}
\]

where \(|X|\) represents the cardinality of a set \(X\).

Similarly, the boundary set can be defined by \((X, U, Ds)\).

When the boundary set is not empty, sample separation axiom does not hold.

### 4 Design Principles of Categorization Methods

If categorization axioms hold for a categorization result, then three design principles of categorization methods can be inferred. In the following, we will discuss such three principles respectively.

#### 4.1 Category Compactness Principle

A categorization result should abide by sample separation axiom. However, sample separation axiom just is a minimum requirement for a categorization result. In theory, it is not sufficient for a good categorization result to solely follow sample separation axiom. Theoretically, a good categorization result should abstain from violating sample separation axiom as much as possible. In other words, it is very important for every object that the intra category similarity should be greater than the inter category similarity as much as possible. As high intra category similarity means low inter category similarity, it is natural for a good categorization result to make the intra category similarity as larger as possible. Obviously, the greater intra category similarity is, the more compact a categorization result becomes. Consequently, when designing a categorization method, category compactness principle can be stated as:

**Category Compactness Principle:** A categorization method should make its categorization result as compact as possible.

More detailed, category compactness principle requires that the maximum similarity between one object and its corresponding category should be as great as possible, which is equivalent to ensuring the minimum dissimilarity between one object and its corresponding category. In other words, compactness, according to the category compactness principle, translates to maximal intra category similarity or minimal intra category variance.

Certainly, the category compactness principle can be directly used to design an categorization criterion when category compactness is defined. Thus, different requirements lead to different definitions of category compactness. A general definition of category compactness criterion can be defined as follows.

**Category Compactness Criterion:** \(J_C : \{I(X)\} \times \{\{X_1, X_2, \ldots, X_c\} | \forall i, X_i \text{ represents } X_i \rightarrow R_i\} \) is called category compactness criterion if the optimum of \(J_C(I(X), \{X_1, X_2, \ldots, X_c\})\) corresponds to the categorization result with the largest category compactness, where \(I(X)\) denotes the input representation of the data set \(X = \{x_1, x_2, \ldots, x_c\}\) without considered the label information.

#### 4.2 Category Separation Principle

If a categorization result \((X, U, Sim)\) satisfies category separation axiom, then \(\forall 1 \leq i \neq j \leq c, X_i \neq X_j\). For a good categorization result, it is not sufficient to satisfy \(\forall 1 \leq i \neq j \leq c, X_i \neq X_j\). Usually, the distance between categories is expected to be as large as possible. Therefore, a categorization criterion should make its categorization result follow cluster separation principle as follows:

**Category Separation Principle:** A good categorization result should have the maximum distance between categories.

Category separation principle means that the distance between categories needs to be defined by a categorization method and the distance between categories should be as large as possible, which follows that the partition outputted by a categorization method should abstain from violating category separation axiom as far as possible.

According to the above analysis, it is very natural that a categorization method makes its categorization result satisfy category separation axiom. The category separation principle requires the development of a categorization criterion for measuring category separation. A general definition of category separation criterion can be defined as follows.

**Category Separation Criterion:**

\(J_S : \{I(X)\} \times \{\{X_1, X_2, \ldots, X_c\} | \forall i, X_i \text{ represents } X_i \rightarrow R_i\} \) is called category separation criterion if the optimum of \(J_S(I(X), \{X_1, X_2, \ldots, X_c\})\) corresponds to the categorization result with maximal category separation.

#### 4.3 Categorization Consistency Principle

If a categorization result \((X, U, Sim)\) satisfies categorization equivalency axiom, then categorization is
consistent. However, even for human recognition systems, categorization equivalency axiom can not be always guaranteed to be true. Generally, human recognition systems always try to make categorization error as small as possible. Therefore, categorization equivalency axiom is a very demanding requirement for categorization. A reasonable categorization criterion should relax categorization equivalency axiom into categorization consistency principle as follows:

Categorization Consistency Principle: A good categorization result should make the categorization equivalency axiom hold as much as possible.

Similarly, categorization consistency principle can be used to design some categorization criterion as follows:

Categorization Consistency Criterion: $J_E : \{U\} \times \{\text{Sim}\} \rightarrow \mathbb{R}^+$ is called categorization consistency criterion if the optimum of $J_E(U, \{\text{Sim}\})$ corresponds to the categorization result which make categorization equivalency axiom hold with the largest probability.

5 Classification and Clustering

When $U$ is known before categorization, categorization is a classification problem. When $U$ is not known, categorization is a clustering problem. The proposed axioms are theoretically available for classification and clustering. In the following, we will discuss them respectively.

5.1 Classification

If $U$ is known before categorization, categorization becomes a classification problem. For classification, a category is called a class. If categorization equivalency axiom is true, the classification error will be zero. In practice, a classification method can only make a classification error as small as possible, but usually its classification error is not zero. Therefore, categorization equivalency axiom should be as a constraint for a classification problem. In other words, when dealing with a classification problem, categorization equivalency axiom should be true as much as possible in probability.

When $U$ is a proper partition, the corresponding classification problem is standard classification problem. When $U$ is a overlapping partition, the corresponding classification problem is multi label classification problem. For multi label classification, sample separation axiom should be generalized as $\forall k \exists i (i \in \arg \max_i \text{Sim}(x_k, X_i))$. Under such a generalization, multi label classification also follows sample separation axiom.

As for category separation axiom, it is usually true for classification. Therefore, categorization equivalency axiom and sample separation axiom are very important for classification. In particular, when $U$ is a proper partition, we can set $\rho(k) = \arg \max_i \text{Sim}(x_k, X_i)$.

Using such denotation, sample separation axiom and categorization equivalency axiom can be rewritten as follows:

Sample Separation Axiom: $\forall x \exists i \forall j ((j \neq i) \rightarrow (\text{Sim}(x, X_i) > \text{Sim}(x, X_j)))$.

Categorization Equivalency Axiom: $\forall x (\rho(x) = h(x))$.

Furthermore, the decision region for a classification result $(X, U, \text{Sim})$ can be defined as follows:

Decision Region:

$\Omega = \{x | \exists i \forall j ((j \neq i) \rightarrow (\text{Sim}(x, X_i) > \text{Sim}(x, X_j)))\}$. In particular, the decision region for a class $X_i$ can be defined as follows:

Decision Region for a Class $X_i$:

$\Omega_i = \{x | \forall j ((j \neq i) \rightarrow (\text{Sim}(x, X_i) > \text{Sim}(x, X_j)))\}$. Therefore, it is easy to know that $\cup_i \Omega_i = \Omega$.

The boundary for a classification result $(X, U, \text{Sim})$ can be defined as follows:

Boundary: $\partial \Omega = \Omega - \Omega^2$, where $\Omega$ represents the closure of $\Omega$, $\Omega^2$ represents the interior of $\Omega$.

The training decision region can be defined as follows:

Training Decision Region: $\Omega(X, U, \text{Sim}) = \{x | \exists i \forall k ((x \in \Omega_i) \land (x_k \in \Omega_i) \land (\text{Sim}(x, X_i) \geq \text{Sim}(x_k, X_i))\}$.

Training Decision Region for a class $X_i$: $\Omega_{X_i} = \{x | \exists k ((x \in \Omega_i) \land (x_k \in \Omega_i) \land (\text{Sim}(x, X_i) \geq \text{Sim}(x_k, X_i)))\}$.

The support vector for a classification result $(X, U, \text{Sim})$ can be defined as follows:

Support Vector: If $x_k \in \partial \Omega(X, U, \text{Sim})$, then $x_k$ is called a support vector for the classification result $(X, U, \text{Sim})$.

The margin for a classification result $(X, U, \text{Sim})$ can be defined as follows:

Margin $(X, U, \text{Sim}) = \min_{i \neq j} d(\Omega_{X_i}, \Omega_{X_j})$, where $d(\Omega_{X_i}, \Omega_{X_j})$ represents the distance between $\Omega_{X_i}$ and $\Omega_{X_j}$.

Categorization equivalency axiom requires that $\forall x (\rho(x) = h(x))$, which is impossible in practice.
as \( \rho(x) \) is not known but only \( \rho(x_k) \) is known for \( k \in \{1, 2, \ldots, n\} \). Therefore, it is natural to relax \( \forall x (\rho(x) = h(x)) \) as \( P(\rho(x) \neq h(x)) \leq \varepsilon \). If \( P(\rho(x) \neq h(x)) \leq \varepsilon \) holds with a probability not less than \( 1 - \delta \), PAC theory has provided a theoretical investigation \((\text{Valiant, 1984})\).

For developing a classification method, categorization consistency principle requires that \( \sum_{k=1}^{n} L(\rho(x_k) - h(x_k)) \) reaches the minimum, which is usually called minimizing empirical risk. Transparently, neural networks can be introduced by minimizing empirical risk. Usually, the more complexity of \( h(x) \), the more small the empirical risk. Therefore, the tradeoff between the empirical risk and the function complexity will lead to the structural risk.

For a probability point of view, categorization consistency principle requires to keep the probability \( P(\rho(x) = h(x)) \) as large as possible, which means that \( \max \sum_{k} \prod_{i} \text{Sim}(x_{ki}, x_{\rho(ki)}) \) is a natural objective function for a classification problem. Transparently, \( \max \sum_{k} \prod_{i} \text{Sim}(x_{ki}, x_{\rho(ki)}) \) abides by category compactness principle. Many classification methods can be followed by this objective function.

For example, K-nearest neighbor classification method defines \( X_{1} = \{x_{k} | \alpha_{ik} = 1 \} \) and \( \text{Sim}(x, X_{j}) = \frac{|N_{j}(x)|}{k} \), \( N_{j}(x) = \{x_{i} | x_{i} \in X_{j} \wedge x_{i} \in \text{K-nearest neighborhood of } x \} \). The categorization result of K-nearest neighbor classification satisfies sample separation axiom.

Let \( \text{Sim}(x, X_{j}) = P(X_{j} | x) \), it is easy to know that the categorization result of Bayesian classification follows sample separation axiom as the input \( x \) belongs to \( X_{i} \) just because \( \text{Sim}(x, X_{i}) = \max_{j} \text{Sim}(x, X_{j}) = \max_{j} P(X_{j} | x) = P(X_{i} | x) \).

Let \( \text{Ds}(x, X_{j}) = R(\alpha_{i} | x) = \sum_{j=1}^{c} \lambda_{ij} P(X_{j} | x) \), where the action \( \alpha_{i} \) denotes the decision to assign the input to class \( X_{i} \) and \( \lambda_{ij} \) denotes the cost incurred for taking the action \( \alpha_{i} \) when the input belongs to \( X_{j} \). Transparently, the categorization result of minimum risk classification abides by sample separation axiom.

Let \( \text{Sim}(x, X_{i}) = U(\alpha_{i} | x) = \sum_{j=1}^{c} U_{ij} P(X_{j} | x) \), where the action \( \alpha_{i} \) denotes the decision to assign the input to class \( X_{i} \) and \( U_{ij} \) measures how good it is to take the action \( \alpha_{i} \) when the input belongs to \( X_{j} \). Maximum expected utility classifier also follows sample separation axiom.

Moreover, let \( g_{i}(x) = \log \text{Sim}(x, X_{i}) \) be discriminant function, sample separation axiom requires that object \( x \) is assigned to class \( X_{i} \) if \( g_{i}(x) = \max_{j} g_{j}(x) \). If \( g_{i}(x) = \log \text{Sim}(x, X_{i}) = w_{i}^{T}x + w_{0} \), linear discrimination analysis also satisfies sample separation axiom.

In particular, when \( c = 2 \), set \( g_{1}(x) = \log \text{Sim}(x, X_{1}) = w_{i}^{T}x + b_{1} \) and \( g_{2}(x) = \log \text{Sim}(x, X_{2}) = -w_{i}^{T}x - b_{1} \). Assuming categorization equivalency axiom holds, it means that \( w_{i}^{T}x + b_{1} \geq 0 \) for \( \forall x_{k} \in X_{1} \) and \( -w_{i}^{T}x - b_{1} \leq 0 \) for \( \forall x_{k} \in X_{2} \). Therefore, category separation principle states that the optimal linear discrimination should keep the distance between the two parallel hyperplanes as large as possible when categorization equivalency axiom holds, which leads to the famous support vector machine.

It is easy to know that the training decision region for support vector machine is \( \Omega_{w} = \{x | w_{i}^{T}x + b_{1} \geq 0 \} \) for \( \forall x_{k} \in X_{1} \) and \( -w_{i}^{T}x - b_{1} \leq 0 \) for \( \forall x_{k} \in X_{2} \). It is easy to prove that \( \text{Margin}_{w} = \frac{1}{\sqrt{w_{i}^{T}w}} \) means a better generalization for support vector machine, which has been proved by statistical learning theory \((\text{Vapnik, 2000})\).

Let \( \text{Ds}(x, X_{1}) = w_{i}^{T}(x - v_{1})(x - v_{1})^{T}w \) and \( \text{Ds}(x, X_{2}) = w_{i}^{T}(x - v_{2})(x - v_{2})^{T}w \), where \( v_{1} = \sum_{x_{k} \in X_{1}} x_{k}, \) and \( v_{2} = \sum_{x_{k} \in X_{2}} x_{k} \). According to category compactness principle, we need to minimize \( w_{i}^{T}S_{w}w = \sum_{x_{k} \in X_{1}} \text{Ds}(x_{k}, X_{1}) + \sum_{x_{k} \in X_{2}} \text{Ds}(x_{k}, X_{2}) \). According to category separation principle, we need to maximize \( w_{i}^{T}S_{w}w = w_{i}^{T}(v_{1} - v_{2})(v_{1} - v_{2})^{T}w \). Combining the above two objective functions, a reasonable objective function is to minimize \( \frac{1}{\sqrt{w_{i}^{T}w}} \), which leads to Fisher linear discriminant analysis. In summary, Fisher linear discriminant analysis is consistent with category compactness principle and category separation principle. Certainly, the categorization result of Fisher linear discriminant analysis follows sample separation axiom.

In summary, classification results of many classification methods follow sample separation axiom. Moreover, many classification methods are developed based on category compactness principle, category separation principle and categorization consistency principle.

### 5.2 Clustering

When \( U \) is not known, categorization becomes a clustering problem. For clustering, a category is called a cluster. In the following, category separation is called cluster separation, category compactness is called cluster compactness, categorization result is called clustering result.

Clearly, proper clustering result and overlapping clustering result are useful in practice. In cluster analysis community, overlapping clustering results are usually considered meaningful \((\text{Jardine and Sibson, 1971})\) and \((\text{Ding and He, 2004})\) and have practical applications, such as categorization in cognitive science \((\text{Shepard and Arabie, 1979})\), community detec-
tion in complex networks [Palla et al., 2005] and movie recommendation [Banerjee et al., 2005], etc. However, a good clustering result is not expected to be an improper clustering result. There are various cases of improper clustering results, as listed in Section 3.

5.2.1 Inequalities on Clustering Results and Some Clustering Criteria

Sample separation axiom follows several inequalities regarding clustering results, such as Theorem 2 and 3.

**Theorem 2.** Let \((X, U, Sim)\) be a clustering result for the given data set \(X = \{x_1, x_2, \cdots, x_n\}\). If sample separation axiom holds, the inequalities (1), (2), (3) and (4) hold.

\[
\prod_k Sim(x_k, X_{\varphi(k)}) \geq \prod_k Sim(x_k, X_{\phi(k)}) \tag{1}
\]

\[
\sum_k Sim(x_k, X_{\varphi(k)}) \geq \sum_k Sim(x_k, X_{\phi(k)}) \tag{2}
\]

\[
\prod_k Sim(x_k, X_{\varphi(k)}) \geq \prod_k \sum_i \alpha_i Sim(x_k, X_i) \tag{3}
\]

\[
\sum_k Sim(x_k, X_{\varphi(k)}) \geq \sum_k f(\sum_i \alpha_i g(Sim(x_k, X_i))) \tag{4}
\]

where \(\varphi(k) = \arg \max_i Sim(x_k, X_i)\), \(\phi(k)\) is a function from \(\{1, 2, \cdots, n\}\) to \(\{1, 2, \cdots, c\}\), \(\alpha_i > 0\), \(\sum_{i=1}^c \alpha_i = 1\); \(f\) is a convex function; and \(\forall t \in R_+\), \(f(g(t)) = t\).

**Theorem 3.** Let \((X, U, Ds)\) be a clustering result for the given data set \(X = \{x_1, x_2, \cdots, x_n\}\). If sample separation axiom holds, the inequalities (5), (6), (7) and (8) hold.

\[
\sum_k Ds(x_k, X_{\varphi(k)}) \leq \sum_k Ds(x_k, X_{\phi(k)}) \tag{5}
\]

\[
\sum_k Ds(x_k, X_{\varphi(k)}) \leq \sum_k f(\sum_i \alpha_i g(Ds(x_k, X_i))) \tag{6}
\]

\[
\prod_k Ds(x_k, X_{\varphi(k)}) \leq \prod_k Ds(x_k, X_{\phi(k)}) \tag{7}
\]

\[
\prod_k Ds(x_k, X_{\varphi(k)}) \leq \prod_k f(\sum_i \alpha_i g(Ds(x_k, X_i))) \tag{8}
\]

where \(\varphi(k) = \arg \min_i Ds(x_k, X_i)\), \(\phi(k)\) is a function from \(\{1, 2, \cdots, n\}\) to \(\{1, 2, \cdots, c\}\), \(\forall t \in R_+\), \(f(g(t)) = t\), \(f\) is a concave function, and \(\alpha_i > 0\) and \(\sum_{i=1}^c \alpha_i = 1\).

Theorem 2 and 3 provides qualitative properties of clustering results when the sample separation axiom holds. Clearly, Theorem 2 and 3 show that the clustering result should reach the optimum of certain functions, which provides a new interpretation of clustering algorithms. For instance, when \(\forall i, X_i\) is fixed, \(\prod_k Sim(x_k, X_{\varphi(k)})\) can be considered as one kind of maximal intra cluster similarity where \(\varphi(k) = \arg \max_i Sim(x_k, X_i)\). By Theorem 2 the right term in inequality (1) can be considered as a cluster compactness criterion for any partition, its maximum can reach the left term in inequality (1). In other words, the right term in inequality (1) can be chosen as a clustering criterion. Interestingly, maximizing such a clustering criterion can directly lead to classification maximum likelihood under assumptions as follows.

Let the objects in cluster \(X_i\) follow the distribution \(p(x, X_i)\), where \(X_i\) is the distribution parameter, \(1 \leq i \leq c\) respectively. Set \(Sim(x_k, X_i) = P(X_i)p(x_k, X_i)\) and \(P(X_i)\) is the mixing proportion such that \(\sum_{i=1}^c P(X_i) = 1\). Theorem 2 requires that the clustering criterion (9) reaches the maximum in order to make the clustering result satisfy sample separation axiom.

\[
CML = \prod_k P(X_i)p(x_k, X_{\phi(k)}), \tag{9}
\]

where \(\phi(k)\) is a function from \(\{1, 2, \cdots, n\}\) to \(\{1, 2, \cdots, c\}\).

Therefore, maximizing \(\prod_k P(X_i)p(x_k, X_{\phi(k)})\) is equivalent to maximizing \(\sum_k \log(P(X_i)p(x_k, X_{\phi(k)}))\), which results in Classification Maximum Likelihood approach for cluster analysis (Celeux and Govaert, 1993).

Similarly, setting \(Sim(x_k, X_i) = P(X_i)p(x_k, X_i)\), we can prove the inequality (10) by performing the similar proof of (5) of Theorem 2.

\[
c^n \prod_k P(X_{\varphi(k)})p(x_k, X_{\varphi(k)}) \geq \prod_k \sum_i P(X_i)p(x_k, X_i) \tag{10}
\]

Obviously, the right term of the inequality (10) is the famous likelihood of mixture model (Redner and Walker, 1984), which shows many model based clustering algorithms follow cluster compactness principle.

Likely, the right term of the inequality (5) in Theorem 3 represent intra cluster variance and can be considered as a clustering criterion. Minimizing such a clustering criterion can directly lead to C-means if setting \(Ds(x_k, X_i) = ||x - v_i||^2\), Theorem 3 requires minimizing of the clustering criterion (11) as follows:

\[
\min_\phi \sum_k ||x_k - v_{\varphi(k)}||^2, \tag{11}
\]

It is easily proved that the optimal \(\phi\) for minimizing the clustering criterion (11) is \(\varphi(k) = arg \min_x ||x_k -
\[ v_i^2. \] Obviously, minimizing of the clustering criterion \( f \) is equivalent to minimizing the clustering criterion of \( C \)-means. By the same method, Theorem 2 also demands to minimize the clustering criterion of General C-means \((Yu, 2005)\) as follows.

\[
\sum_k f\left(\sum_i \alpha_i g(\|x_k - v_i\|^2)\right),
\]

(12)

In addition, Theorem 2 and 3 can lead to some new clustering criteria. In the following, two new clustering algorithms are presented based on Theorem 2 in order to show that the proposed axioms are helpful to design new clustering algorithms.

Let \( f = x^m \) where \( m \geq 1 \), the right term of inequality (14) in Theorem 2 can be used as the objective function (13) of a new clustering algorithm as follows:

\[
\sum_k \left(\sum_i \alpha_i \text{Sim}(x_k, X_i)\right)^m
\]

(13)

In theory, maximizing (13) can lead to new different clustering algorithms depending on how to define \( \text{Sim}(x_k, X_i) \).

Assume the data set \( X \) can be represented by a feature matrix. Set \( \text{Sim}(x_k, X_i) = \exp(-\beta^{-1}\|x_k - v_i\|^2) \), then (13) becomes (14).

\[
\sum_k \left(\sum_i \alpha_i \exp(-m\beta^{-1}\|x_k - v_i\|^2)\right)^m
\]

(14)

Assume the data set \( X \) can be represented by an adjacency matrix \( A = [A_{kl}]_{n \times m} \). Set \( \text{Sim}(x_k, X_i) = \prod_i \theta^{s_{ik}}_{ui} \), where \( \sum_i \theta_{ui} = 1 \), then (13) becomes (15).

\[
\sum_k \left(\sum_i \alpha_i \prod_i \theta^{s_{ik}}_{ui}\right)^m
\]

(15)

5.2.2 On Clustering Algorithms

In the literature, many authors tried to define cluster by internal cohesion (homogeneity) and external isolation (separation), as Everitt et al. (2011) have declared. For example, Jain (2010) said that an ideal cluster can be defined as a set of points that is compact and isolated, more detailed Bonner (1964) defined tight clusters by maximal complete subgraphs of the similarity matrix graph, Abell (1955) defined compact cluster as at least 50 members are within a given radial distance of the cluster center. Transparently, the idea of cluster compactness and cluster separation has been used to develop clustering algorithm in the literature, but most definitions about compactness or separation are originated from intuition and applications. For example, objective functions of many existed clustering algorithms are consistent with cluster compactness principle, just name a few, K-modes (Huang, 1998), General stochastic block model (Shen et al., 2011), CNM (Clauset et al., 2004), pairwise data clustering (Hofmann and Buhmann, 1997), Nonnegative Matrix Factorization (Xu et al., 2003), dominant set clustering (Pavan and Pelillo, 2007), and so on; several cluster separation criteria have been proposed based on cluster separation principle such as Girvan and Newman’s algorithm \((GN\) algorithm\) (Girvan and Newman, 2002), ratio cut (Wei and Cheng, 1989), normalized cut (Shi and Malik, 2000), Maximum Margin Clustering (Ben-Hur et al., 2002), and so on.

In this paper, cluster compactness principle and cluster separation principle have been clearly followed by separation axioms. Moreover, a proper clustering result satisfies the categorization equivalency axiom. Therefore, the categorization equivalency axiom can be directly used as a principle to design a clustering algorithm, which is also called categorization consistency principle.

According to the categorization consistency principle, a cognitive representation is equivalent to an extensional representation for a clustering result with respect to Sim or \( Ds \). For a given clustering result, its extensional representation of can be generated by its cognitive representation with cluster similarity mapping. Inversely, the cognitive representation can be determined by the corresponding extensional representation and the given data set. Therefore, a clustering result should be a convergence point for the above-mentioned iterative process. If the extensional representation is considered a partition, a common method for developing a clustering algorithm is to alternately update the partition and the cognitive representation, such as in Runkler and Bezdek (1993), MacQueen et al. (1967). When such an iterative process converges, it is easy to observe that the categorization equivalency axiom holds for the outputted clustering result. Hence, a framework for an iterative clustering algorithm is developed based on the categorization equivalency axiom, as demonstrated in Algorithm 1.

In the literature, many clustering algorithms have been described by Algorithm 1. For example, set \( Ds(x_k, X_i) = \|x_k - v_i\|^2 \), \( X_i = v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}} \), \( u_{ik} = 1 \) if \( i = \arg \min_j \|x_k - v_j\|^2 \) otherwise \( u_{ik} = 0 \), Algorithm 1 results in C-means.

Set \( \text{Sim}(x_k, X_i) = \max_{x_j \in X_k} \text{Sim}(x_k, x_j) \) and \( u_{ik} = 1 \) if \( i = \arg \max_j \text{Sim}(x_k, X_i) \) otherwise \( u_{ik} = 0 \), Algorithm 1 is Single Linkage, \( X_i = X_i \).

set \( Ds(x_k, X_i) = \|x_k - v_i\|^2 \), \( X_i = v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}} \), \( u_{ik} = \|x_k - v_i\|^2 / \sum_{i=1}^{c} \|x_k - v_i\|^2 \), Algorithm 1
Algorithm 1 Framework of Iterative Clustering Algorithm.

Require:
- $I(X)$ represents the data set $X$,
- Initial Partition $U$;
- Convergence Thresholds;
- Free Parameters;

1: repeat
2: Update $X_1, X_2, \ldots, X_c$ by a partition $U$;
3: Update $U$ by cluster similarity mapping $Sim(x_k, X_i)$;
4: until Convergence thresholds are satisfied, output $U$ or $\{X_1, X_2, \ldots, X_c\}$

results in fuzzy C-means.
Set $Sim(x_k, X_i) = P(X_i)p(x_k|X_i)$, $u_{ik} = P(X_i|x_k)$ and $P(X_i) = \sum_{k=1}^c u_{ik}/n$, Algorithm 1 leads to model based clustering.

When the boundary set is empty, Algorithm 1 outputs a proper clustering result.

5.2.3 Separation Axioms and Cluster Validity

According to Liu et al. (2013), there are two widely accepted methods when designing a cluster validity index: one is to measure how distinct or well-separated a cluster is from other clusters (which is called the separation criterion), consistent with our proposed category separation axiom; the other is to measure how compact a cluster itself is (which is called the compactness criterion), consistent with sample separation axiom. These two criteria are consistent with the cluster separation principle and cluster compactness principle.

Furthemore, a clustering result that violates separation axioms cannot be considered to have the best clustering quality. Consequently, clustering results similar to such improper clustering results should be regarded as being not as good as those far from the improper clustering results. Adopting such a view is useful in designing a cluster validity index, which leads to another principle for designing a cluster validity index based on the separation axioms:

**Extreme Value Principle:** A good cluster validity index should evaluate improper clustering results as those exhibiting the poorest clustering quality.

In the literature, there are many cluster validity indices that are only consistent with one or two of the three above-mentioned principles. For instance, the partition coefficient $V_{pe} = \frac{1}{n} \sum_k \sum_{i} u_{ik}^2$ (Bezdek, 1974) and partition entropy $V_{pe} = \frac{1}{n} \sum_i \sum_k u_{ik} \log u_{ik}$ (Bezdek, 1975) only reflect the cluster compactness principle and the extreme value principle. Fuzzy hypervolume validity (Gath and Geva, 1989) only reflects the cluster compactness principle. However, different from clustering criterion design, many cluster validity indices in Table 1 are usually developed based on the above three above-mentioned principles.

As an example, we will show that Xie-Beni Index is indeed consistent with the proposed principles of designing a cluster validity index (Xie and Beni, 1991). Xie-Beni Index is a well known cluster validity index for fuzzy C-means, and it is defined as (16).

$$XB(X, U, V) = \sum_{i=1}^c \sum_{k=1}^n \frac{u_{ik}^2 \|x_k - v_i\|^2}{n \times \min_{i \neq j} \|v_i - v_j\|^2}$$

The larger $XB(X, U, V)$ is, the worse the clustering result $(U, V)$ becomes. The smaller $XB(X, U, V)$ is, the better the clustering result $(U, V)$. Clearly, coincident partition will cause $XB(X, U, V)$ to approach infinity and should thus be considered an improper clustering result. In $XB(X, U, V)$, the numerator represents the compactness of the partition, the denominator represents the separation degree of the partition. Obviously, $XB(X, U, V)$ simultaneously considers cluster compactness, cluster separation and extreme value principle. Xie-Beni index was not originated from separation axioms but is naturally consistent with our proposed separation axioms. It should be noted that many recently developed cluster validity indices are consistent with the proposed principles; for example, CVNN (Liu et al., 2013) considers both the cluster compactness and cluster separation principles.

6 Discussion and Conclusions

In this paper, categorization results are represented by $(X, U, Sim)$ or $(X, U, Ds)$ based on cognitive science in a united way. Based on such the proposed representation of categorization result, three categorization axioms are proposed for a categorization result. When $U$ is a hard partition, prototype theory of categorization and exemplar theory of categorization are consistent with sample separation axiom. The reason is very simple, prototype theory of categorization (Rosch, 1978) states that an object is categorized as category A not other categories just because it is more similar to the prototype of category A than it is to the prototypes of other categories; exemplar theory of categorization (Medin and Schaffer, 1978) states that an object is categorized as category A not other categories just because it is more similar to multiple exemplars of category A than it is to those multiple exemplars of other categories. With respect to sample separation axiom, the difference between prototype theory and
Table 1: Several cluster validity indices based on separation axioms

| Index                      | Optimal number of clusters |
|----------------------------|-----------------------------|
| $K_{won}$ = $\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{u_{ik}^2 \|x_k - v_i\|^2 + c - 1}{\sum_{i=1}^{n} \sum_{j=1}^{n} \min(u_{ik}, u_{ij})}$ | Minimum                      |
| $V_P = \frac{1}{n} \sum_{i=1}^{n} \max(u_{ik}) - \frac{2}{(c-1)} \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n} \min(u_{ik}, u_{ij})$ | Maximum                      |
| $DB = \frac{1}{nc} \sum_{k \neq c} \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n} \min(u_{ik}, u_{ij})$ | Minimum                      |
| $FS = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik} \|x_k - v_i\|^2 - \sum_{k=1}^{n} u_{ik} \|\bar{v} - v_i\|^2$ | Minimum                      |
| $Silhouette = \frac{1}{nc} \sum_{k=1}^{c} \left( \frac{\sum_{i=1}^{n} d(x_k, v_i)}{n_k \times \max(b(x_k), a(x_k))} \right)$ | Maximum                      |
| $I = \frac{1}{nc} \sum_{k=1}^{c} \sum_{i=1}^{n} \frac{n_k (b(x_k) - a(x_k))}{\max(b(x_k), a(x_k))}$ | Maximum                      |
| $Dunn = \min_{i=1}^{n} \min_{j=1}^{n} \frac{d(x_i, x_j)}{\sum_{k=1}^{c} \sum_{x_k \in X_i} d(x_k, x_j)}$ | Maximum                      |
| $CH(X, v, c) = \frac{(n-c) \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^2 \|v_i - v_c\|^2}{(c-1) \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij}^2 \|v_i - v_j\|^2}$ | Maximum                      |

Exemplar theory is category representation. Clearly, separation axioms present the common properties that categorization results should satisfy. Categorization equivalency axiom is an ideal case for a categorization result, which is consistent with a basic assumption: the category in the mind fits the category in the world.

Separation axioms not only classify categorization results into proper categorization result, overlapping categorization result and improper categorization result, but also classify partitions into proper partition, overlapping partition and improper partition. Moreover, categorization axioms can lead to three design principles of categorization methods: category compactness principle, category separation principle and categorization consistency principle.

When $U$ is known, categorization becomes classification. As for classification, category separation axiom is always true for a classification result, but sample separation axiom is true for a proper classification result and categorization equivalency axiom just holds for a classification result as much as possible. Therefore, sample separation axiom and categorization equivalency axiom are more important constraints for classification. With respect to a classification result ($X, U, Sim$), decision region, training decision region and margin are defined by sample separation axiom. As categorization methods, category compactness principle can result in K-nearest classification, Bayesian classification, Minimum risk classification, Maximum expected utility classification, Linear discriminant analysis. Category separation principle can lead to support vector machine. Combining category compactness principle and category separation principle can introduce Fisher linear discriminant analysis. Categorization consistency principle can lead to empirical risk and structural risk, which can result in neural networks.

When $U$ is not known, categorization becomes clustering. As for clustering, sample separation axiom, cluster separation axiom and categorization equivalency axiom are the same constraints for a proper clustering result. Several inequalities regarding clustering results are also obtained based on sample separation axiom. Categorization equivalency axiom can lead to a framework for an iterative clustering algorithm. Moreover, extreme value principle is proposed for designing a cluster validity index except cluster separation principle and cluster compactness principle.

In the future, we need further study boundary and training decision region for classification, in particular, the relation between the margin and the generalization ability for a classification result. For SVM, the largest margin means a better generalization. A natural question is whether or not a classification result with the largest margin have the best generalization ability. Second, we need to investigate whether the proposed axioms can introduce more interesting properties about categorization results, and develop some new categorization methods. Third, we hope to further study overlapping categorization, including multi-label classification and overlapping clustering. Fourth, we hope to investigate how to extend categorization axioms into regression and reinforcement learning. In the final, improper categorization needs to be further...
Acknowledgements

We thank Xinbo Gao, Wensheng Zhang, Baogang Hu, Jiangshe Zhang, Liping Jing and Bianfang Chai, their valuable discussions and suggestions have greatly improved the presentation of this paper. This work was supported by the NSFC grants (61033013, 61370129), Ph.D Programs Foundation of Ministry of Education of China (2012009110006), PCSIRT (IRT 201206), Beijing Committee of Science and Technology, China(Grant No. Z131110002813118).

References

Abell, G. O. (1958). The distribution of rich clusters of galaxies. The Astrophysical Journal Supplement Series, 3:211.

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, R. J. (2005). Model-based overlapping clustering. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 532–537. ACM.

Basu, S. et al. (2005). Semi-supervised clustering: probabilistic models, algorithms and experiments. University of Texas at Austin.

Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2002). Support vector clustering. The Journal of Machine Learning Research, 2:125–137.

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. Journal of Cybernetics, 3:58–72.

Bezdek, J. C. (1975). Mathematical models for systematics and taxonomy. In Proceedings of eighth international conference on numerical taxonomy, volume 3, pages 143–166.

Bonner, R. E. (1964). On some clustering techniques. IBM journal of research and development, 8(1):22–32.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1):1–27.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. Journal of Statistical Computation and simulation, 47(3-4):127–146.

Chen, M.-Y. and Linkens, D. A. (2001). Rule-base self-generation and simplification for data-driven fuzzy models. In The 10th IEEE International Conference on Fuzzy Systems, 2001., volume 1, pages 424–427. IEEE.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6):066111.

Correa-Morris, J. (2013). An indication of unification for different clustering approaches. Pattern Recognition.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2):224–227.

Ding, C. and He, X. (2004). K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In Proceedings of the 2004 ACM symposium on Applied computing, pages 584–589. ACM.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1):95–104.

Everitt, B. S., Landau, S., Morven, L., and Stahl, D. (2011). Cluster Analysis. John Wiley & Sons, Ltd.

Fukuyama, Y. and Sugeno, M. (1989). A new method of choosing the number of clusters for the fuzzy c-means method. In Proc. 5th Fuzzy Syst. Symp, volume 247.

Gath, I. and Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7):773–780.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826.

Hofmann, T. and Buhmann, J. M. (1997). Pairwise data clustering by deterministic annealing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(1):1–14.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3):283–304.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8):651–666.

Jardine, N. and Sibson, R. (1971). Mathematical taxonomy. London etc.: John Wiley.

Kleinberg, J. (2003). An impossibility theorem for clustering. Advances in neural information processing systems, pages 463–470.

Kwon, S. H. (1998). Cluster validity index for fuzzy clustering. Electronics Letters, 34(22):2176–2177.

Lakoff, G. (1987). Women, Fire, and Dangerous Things. University of Chicago Press, Chicago.
Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., and Wu, S. (2013). Understanding and enhancement of internet clustering validation measures. *IEEE Transactions on Cybernetics*, 43(3):982–994.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. California, USA.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207.

Murphy, G. L. (2004). *The big book of concepts*. MIT press.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3):289.

Palla, G., Derényi, I., Farkas, L., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Pavan, M. and Pelillo, M. (2007). Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):167–172.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Roch, E. (1978). Principles of categorization. *Cognition and categorization*, pages 27–48.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Runkler, T. A. and Bezdek, J. C. (1999). Alternating cluster estimation: A new tool for clustering and function approximation. *IEEE Transactions on Fuzzy Systems*, 7(4):377–393.

Shen, H.-W., Cheng, X.-Q., and Guo, J.-F. (2011). Exploring the structural regularities in networks. *Physical Review E*, 84(5):056111.

Shepard, R. N. and Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.

Wei, Y.-C. and Cheng, C.-K. (1989). Towards efficient hierarchical designs by ratio cut partitioning. In *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers*, 1989 IEEE International Conference on, pages 298–301. IEEE.

Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–273. ACM.

Yu, J. (2005). General c-means clustering model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1197–1211.

### A The Proof of Theorem 1

**Proof.** For a categorization result \((\bar{X}, U, Sim)\), there exist \(i \neq j\) such that \(X_i = X_j\). According to category separation axiom, for category \(X_i\), there exists an object \(x_k\) such that \(Sim(x_k, X_i) > Sim(x_k, X_j)\). However, \(X_i = X_j\) means that \(Sim(x_k, X_i) = Sim(x_k, X_j)\). It is a contradiction. In other words, the first conclusion is proved.

Similarly, if category separation axiom hold, for category \(X_i\) there exists an object \(x_k\) such that \(\forall j(j \neq i) \rightarrow (Sim(x_k, X_i) > Sim(x_k, X_j))\). If there exist \(i \neq j\) such that \(x_{k_i} = x_{k_j}\), then \(Sim(x_{k_i}, X_i) > Sim(x_{k_i}, X_j)\) and \(Sim(x_{k_j}, X_i) > Sim(x_{k_j}, X_j)\). Since \(x_{k_i} = x_{k_j}\), it means that \(Sim(x_{k_i}, X_i) > Sim(x_{k_j}, X_j)\) and \(Sim(x_{k_j}, X_i) > Sim(x_{k_i}, X_j)\), which is a contradiction. Therefore, the second conclusion is proved. Hence, the proof is finished. \(\square\)

### B The Proof of Theorem 2

**Proof.** 1) Because the sample separation axiom holds, \(\varphi(k) = \arg \max_i Sim(x_k, X_i)\) is well defined and the inequality \(Sim(x_k, X_{\varphi(k)}) \geq Sim(x_k, X_{\varphi(k)}) \geq 0\) must hold. Therefore, inequality (1) can be easily proven by multiplying the above inequality according to subscript \(k\) from 1 to \(n\).

2) Similarly, inequality (2) can be proven.
3) Because $\text{Sim}(x_k, X_{\varphi(k)}) \geq \text{Sim}(x_k, X_i) \geq 0, \forall i$, therefore, we know that inequality (17) holds.

$$\alpha_i \text{Sim}(x_k, X_{\varphi(k)}) \geq \alpha_i \text{Sim}(x_k, X_i) \geq 0.$$ (17)

By summing inequality (17) according to subscript $i$ from 1 to $c$, we can obtain inequality (18).

$$\sum_i \alpha_i \text{Sim}(x_k, X_{\varphi(k)}) \geq \sum_i \alpha_i \text{Sim}(x_k, X_i) \geq 0.$$ (18)

Because $\sum_{i=1}^c \alpha_i = 1$, the inequality (3) can be easily proven by multiplying the inequality (18) according to subscript $k$ from 1 to $n$.

4) Because $f$ is a convex function, we know that inequality (19) holds.

$$\sum_i \alpha_i f(g(\text{Sim}(x_k, X_i))) \geq f(\sum_i \alpha_i g(\text{Sim}(x_k, X_i)))$$ (19)

Because $\forall t \in R_+, f(t) = t$, we know inequality (19) becomes inequality (20).

$$\sum_i \alpha_i \text{Sim}(x_k, X_i) \geq f(\sum_i \alpha_i g(\text{Sim}(x_k, X_i)))$$ (20)

By performing an analysis similar to that described above, we can obtain inequality (4).

Thus, the proof of Theorem 2 is complete. □

C The Proof of Theorem 3

By performing similar analysis of Theorem 2 we can prove Theorem 3.