2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

Clustering Data and Imprecise Concepts

Weifeng Zhang
Intelligent Computing and Machine Learning Lab
School of Automation Science and Electrical Engineering
Beihang University, China
Email: zwf.zhang@gmail.com

Zengchang Qin
[1] Intelligent Computing and Machine Learning Lab
School of ASEE, Beihang University
[2] Robotics Institute, CMU, USA
Email: zcqin@andrew.cmu.edu

***Abstract*—Cluster analysis is the assignment of grouping a set of observations into clusters so that observations in the same cluster are similar in some sense. One of the key features for clustering is how to define a sensible similarity measure. However, classical clustering algorithms have no ability to cluster data instances and imprecise concepts using traditional distance measures. In this paper, we proposed a (dis)similarity measure based on a new knowledge representation framework called label semantics. Based on this new measure, we can automatically cluster data instance and descriptive concepts represented by logical expressions of linguistic labels. Experimental results on a toy problem in image classification demonstrate the effectiveness of the new proposed clustering algorithm. Since the new proposed measure can be extended to measuring distance between any two granularities, the new clustering algorithms can also be extended to clustering data instance and imprecise concepts represented by other granularities.**

***Index Terms*—Clustering; Label Semantics; Linguistic Expressions; K-means; Imprecise Concept Modeling**

I. Introduction

Cluster analysis is considered as the most important forms of unsupervised learning [17]. It deals with finding similar patterns in a collection of unlabeled data. A cluster is therefore a collection of objects which are "similar" and are "dissimilar" to the objects belonging to other clusters. Another kind of clustering is conceptual clustering. Two or more objects belong to the same cluster if this one defines a concept common to all objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Clustering algorithms can be applied in many fields, such as marketing, bio-informatics, video analysis [13] and so on. Hence, in past decades, many classical clustering algorithms, for example K-means, Fuzzy K-means, Hierarchical clustering algorithms, have been proposed and some of them have been successfully applied in resolving practical problems. They group the "similar" objects into one cluster. In this sense, the clustering results depend heavily on the distance measure between objects. Euclidean distance and Mahalanobis distance are two of the most used distance measure. Other (dis)similarity measures may be used for some particular data type. For example, Kullback-Leibler (KL) distance is good for measuring the divergence between two probability distributions. But its unsymmetrical properties make it unsuitable as to be a good universal measure of data distance. In literatures, the objects handled by these classical clustering algorithms are restricted to numerical data, though they could be high-dimensional complex data. However, what we hope to cluster are not limited to numerical data, it could be some high-level knowledge of imprecise concepts or linguistic descriptions [11]. For example, if we are given a set of human age $Age = \{16, 17, 19, 22, 40, 67, 68, 70, 82\}$ and three descriptive concepts $\{young, middle\text{-}aged, old\}$ which are defined by fuzzy membership functions. How can we cluster these two type of data if given the cluster number $K = 3$? Ideally, if the imprecise concept of human age are sensibly defined, we should obtain the following 3 clusters: $\{young, 16, 17, 19, 22\}$, $\{middle\text{-}aged, 40\}$ and $\{old, 67, 68, 70, 82\}$.

Unfortunately, these objects can not be well handled by the existed clustering algorithms. The main reason is there is no good metric for measuring dissimilarity between numerical data and linguistic descriptions [3][6]. In this paper we proposed a novel distance measure which can measure the distance between numeral data and logical expressions of linguistic labels based on *label semantics* [5]. Label semantics uses a set of linguistic labels to represents imprecise concepts in terms of a probability distribution on appropriate label sets. This new proposed measure make it possible to cluster a set of objects including numerical data, concepts, and linguistic descriptions, by modifying the distance measure based on the classical K-means clustering algorithm.

II. Label Semantics

In order to understand the use of natural language for information and knowledge processing in computer systems, Zadeh proposed the concept of Computing with Words paradigm [14]. He suggested a form of precisiated natural language (PNL) based on the theory of generalized constraints and linguistic variables: $\mathbf{x}$ *isr* θ, where $\mathbf{x}$ is the constrained variable, r is the constraining relation, and r is a discrete valued modal variable. The "is" in *isr* is simply its natural meaning- the conjugated verb "to be" and other defined modalities include: possibilistic ($r = blank$); probabilistic ($r = p$); veristic ($r = v$); random set ($r = rs$); fuzzy graph($r = fg$); bimodal ($r = bm$); and Pawlakset ($r = ps$) [15]. However, since the invention of PNL, there is no applicable calculus developed to solve problems in natural language understanding, though some thoughts had been used in designing a search engines by combining it with other computational linguistics technologies [1][10]. Label semantics, proposed by Lawry [4] [5], provides an alternative

978-1-4244-7317-5/11/$26.00 ©2011 IEEE

representation for modeling with words. In contrast to Zadeh's methodology this approach is based on measures of an agent's subjective belief that a logical expression is appropriate to describe a particular object or value. It provides a set of applicable calculus that has been successfully used in many supervised learning applications [4] [9] [11].

Label semantics is a random set framework for modeling with words. The fundamental notion underlying label semantics is that when individuals make assertion of the form ' $\mathbf{x}$ is θ ', different from Zadeh's approach discussed above [15], they are essentially providing information about what labels are appropriate for the value of the underlying variable $\mathbf{x}$ [5]. For $x \in \Omega$ the label description of x is a random set from V into the power set of $\mathbb{L}$, denoted by D_x, with associated distribution m_x, which is referred to as mass assignment:

$$\forall S \subseteq \mathbb{L}, m_x(S) = P(I \in V : D_x^I = S). \tag{1}$$

To evaluate how appropriate a label is for describing a particular value of variable x, *appropriateness degrees* are defined.

$$\forall x \in \Omega, \forall L \in \mathbb{L}, \mu_L(x) = \sum_{S \subseteq \mathbb{L}: L \in S} m_x(S) \tag{2}$$

Example 1: Given a set of labels defined on the height of an adult: $\mathbb{L}_{Height} = \{short, medium, tall\}$ and a voting space with size of 10. Suppose 4 of 10 people agree that *medium* is the only appropriate label for the height of $173cm$ and 6 support that both *short* and *medium* are appropriate labels, according to Eq. 1, the mass assignment for $173cm$ is: $m_{173} = \{medium\} : 0.4, \{short, medium\} : 0.6$. Based on the appropriateness measure given in Eq. 2, the appropriateness degree of *medium* as a description of $173cm$ is

$$\mu_{medium}(173) = 0.4 + 0.6 = 1 \tag{3}$$

and that of *short* is $\mu_{short}(173) = 0.6$.

The mass assignment is extendable to multi-dimensional case. For example, multi-dimensional color labels $\{red, orange, yellow, \ldots\}$ can be used to describe a pixel of color image $\mathbf{p}$ in the HSV space. The mass assignment can be represented in the form of

$$m_{\mathbf{p}} = \{red, orage\} : 0.3, \{red\} : 0.7$$

Such representations will be used in our experiments.

Given a universe of discourse Ω containing a set of objects or instances to be described, it is assumed that all relevant expression can be generated recursively from a finite set of basic labels $\mathbb{L} = \{L_1, L_2, ..., L_n\}$ Operators for combing expressions are restricted to the standard logical connectives of negation '$\neg$', conjunction '$\wedge$', disjunction '$\vee$' and implication '$\rightarrow$'. Hence, the set of logical expressions of labels can be formally defined as follows:

Definition 1 (Logical expressions of labels): The set of logical expressions, *LE*, is defined recursively as follows:

(i) $L_i \in LE \quad for \quad i = 1, 2, ..., n$. (ii) If θ, $\varphi \in LE$ then $\neg\theta, \theta \wedge \varphi, \theta \vee \varphi, \theta \rightarrow \varphi \in LE$.

Basically, we interpret the main logical connectives as follows:

- $\neg L$ means that L is not an appropriate label.
- $L_1 \wedge L_2$ means both L_1 and L_2 are appropriate labels.
- $L_1 \vee L_2$ means that either L_1 or L_2 are appropriate labels.
- $L_1 \rightarrow L_2$ means that L_2 is an appropriate label whenever L_1 is.

As well as labels for a single variable, we may want to evaluate the appropriateness degree of a complex logical expression $\theta \in LE$. Consider the set of logical expressions LE obtained by recursive application of the standard logical connectives. In order to evaluate the appropriateness degrees of such expressions we must identify what information they provide regarding the appropriateness of labels. In general, for any logical expression θ we should be able to identify a maximal set of label sets, $\lambda(\theta)$, that are consistent with θ so that the meaning of θ can be interpreted as the constraint $D_x \in \lambda(\theta)$.

Definition 2 (λ-Function): Let θ and φ be expressions generated by recursive application of the connectives $\neg, \wedge, \vee$ and $\rightarrow$ to the elements of $\mathbb{L}$ (i.e. $\theta \in LE$). Then the set of possible label sets defined by a linguistic expression can be determined recursively as follows:

(i) $\lambda(L_i(x)) = \{S \subseteq \mathbb{F} | \{L_i\} \subseteq S\}$.
(ii) $\lambda(\neg\theta) = \overline{\lambda(\theta)}$.
(iii) $\lambda(\theta \vee \varphi) = \lambda(\theta) \bigcap \lambda(\varphi)$.
(iv) $\lambda(\theta \wedge \varphi) = \lambda(\theta) \bigcup \lambda(\varphi)$.
(v) $\lambda(\theta \rightarrow \varphi) = \overline{\lambda(\theta)} \bigcap \lambda(\varphi)$.

where $\mathbb{F}$, which is refereed to as *focal set*, represents the subsets of labels with possible non-zero masses into the power set of $\mathbb{L}$. The formal definition of focal set is as follows:

Definition 3 (Focal set): Given a universe Ω for variable x, the focal set of $\mathbb{L}$ is a set of focal elements defined as:

$$\mathbb{F} = \{S \subseteq \mathbb{L} | \exists x \in \Omega, m_x(S) > 0\} \tag{4}$$

In label semantics theory, λ-function is a useful tool to represent logical expressions into a set of linguistic labels [11], i.e.: $\lambda(\mathbb{L}) \rightarrow \mathbb{F}$. More details on the label semantics theory are available in [4]. Label semantics has been well applied in data mining and machine learning, a brief review on label semantics based data mining algorithms are given in [12].

III. Distance of Logical Expressions

Based on theory of fuzzy sets many similarity/dissimilarity measures [3] [6] have been proposed for measuring the degree of similarity between fuzzy sets. But those measures are not proper to deal with the similarity/dissimilarity measures between logical expressions which are concepts based on given linguistic variables. Label semantics focus on the decision making process, an intelligent agent must go through in order to identify which labels or logical expressions can actually be used to describe an object or value. For this reason, the appropriateness degree is proposed for measuring the appropriateness of using a particular subset of labels to describe

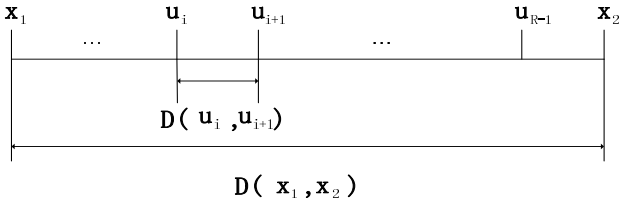

Fig. 1. Illustration of calculating the distance between two data points. The overall dissimilarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ is the aggregated dissimilarity of all the neighboring points $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$ for $i \in R$.

an object or value. One step further, we present a measure to evaluate the dissimilarity of logical expressions based on the mass assignments which can quantize the divergences between logical expressions. This measure is also extendable to measure distance between any two granular sets.

Definition 4 (Distance between data points): Given two data points $\mathbf{x}_1$ and $\mathbf{x}_2$ from an multi-dimensional universe Ω who is fully covered by n labels $\mathbb{L} = \{L_1, ..., L_n\}$, then the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ in this linguistic label space is defined by:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=0}^{R-1} D(\mathbf{u}_i, \mathbf{u}_{i+1}) \tag{5}$$

where

$$D(\mathbf{u}_i, \mathbf{u}_{i+1}) = \sum_{S \in \lambda(\mathbb{L})} (m_{\mathbf{u_i}}(S) - m_{\mathbf{u_{i+1}}}(S))^2 \tag{6}$$

$$\mathbf{u}_i = \mathbf{x}_1 + \frac{i}{R}(\mathbf{x}_2 - \mathbf{x}_1) \tag{7}$$

Fig. 1 intuitively explains the above definition. To calculate the distance between two data points, we divide the distance between them into R ($R > 0$) pieces and calculate the the dissimilarity $D(\mathbf{u}_i, \mathbf{u}_{i+1})$ between two neighboring points $(\mathbf{u}_i, \mathbf{u}_{i+1})$ in the space of linguistic labels. The overall dissimilarity $D(\mathbf{x}_1, \mathbf{x}_2)$ is the sum of the above R pieces of dissimilarity. Thus the accuracy of the distance measure will be improved with increasing the value of R. As we can easily seen that this measure can be generalized to calculate the dissimilarity between any granular sets or a data point to a granular set. In addition, this distance measure has two important properties:

Theorem 1: The distance defined by Equation 5 is symmetric.

$$D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i) \tag{8}$$

Theorem 2: If given $\mathbf{x}_1 < \mathbf{x}_2 < \mathbf{x}_3$[1], then

$$D(\mathbf{x}_1, \mathbf{x}_3) \geq D(\mathbf{x}_1, \mathbf{x}_2) \tag{9}$$

Theorem 2 demonstrates that this measure has non-negative correlation with the distance of $\mathbf{x}$. Both theorems have been proved in [16].

Above definition tells us the way of calculating distance between two data elements. Now we consider the question of how can we design a measure to calculate the distance between a data element point and an imprecise concept represented by logical expressions of linguistic labels. Given a data point $\mathbf{x}_0$ and a linguistic label set $S \in \mathbb{F}$ that covers a continuous area $\delta(S)$ on the universe Ω, the distance between $\mathbf{x_0}$ and S is defined as follows:

$$D(\mathbf{x}_0, S) = \frac{\int_{\delta(S)} D(\mathbf{x}_0, \mathbf{x}) d\mathbf{x}}{\delta(S)} \tag{10}$$

Furthermore, distance of sets of labels which is used to measure the divergence between them can be defined as follows:

Definition 5 (Distance between sets of labels): Given two sets of labels S_i, $S_j \in \mathbb{F}$ and S_i covering a continuous area $\delta(S_i)$ and S_j covering a continuous area $\delta(S_j)$. Then distance between these two sets is defined as:

$$D(S_i, S_j) = \frac{\int_{\delta(S_i)} \int_{\delta(S_j)} D(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j}{\delta(S_i)\delta(S_j)} \tag{11}$$

where $\mathbf{x}_i \in \delta(S_i), \mathbf{x}_j \in \delta(S_j)$, and when $i = j$:

$$D(S_i, S_j) = D(S_i, S_i) = 0 \tag{12}$$

For the symmetry of the distance between two variables, it is obviously that the distance of sets of labels is also symmetric. As we have discussed in Section II, λ-function provides a way of mapping from logical expressions of labels to random set descriptions of labels. So we also can define the distance between two logical expressions as the following.

Definition 6 (Distance between logical expressions): Given two logical expressions $\theta, \varphi \in LE$, then the distance between θ and φ is

$$D(\theta, \varphi) = \frac{1}{pr} \sum_{i=1}^{r} \sum_{j=1}^{p} D(S_i^{\theta \wedge \neg \varphi}, S_j^{\varphi}) - \frac{1}{qt} \sum_{k=1}^{t} \sum_{l=1}^{q} D(S_k^{\varphi \wedge \neg \theta}, S_l^{\theta}) \tag{13}$$

where p, q, r, t, respectively, represent the cardinality of label set $\mathbb{S}^{\varphi}$, $\mathbb{S}^{\theta}$, $\mathbb{S}^{\theta \wedge \neg \varphi}$, $\mathbb{S}^{\varphi \wedge \neg \theta}$, respectively, where:

$$S_i^{\theta} \in \mathbb{S}^{\theta} = \{S | S \in \lambda(\theta)\}, i = 1, 2, ..., q.$$

$$S_j^{\varphi} \in \mathbb{S}^{\varphi} = \{S | S \in \lambda(\varphi)\}, j = 1, 2, ..., p.$$

$$S_k^{\theta \wedge \neg \varphi} \in \mathbb{S}^{\theta \wedge \neg \varphi} = \{S | S \in \lambda(\theta) \bigcap \overline{\lambda(\varphi)}\}, k = 1, 2, ..., r.$$

$$S_l^{\varphi \wedge \neg \theta} \in \mathbb{S}^{\varphi \wedge \neg \theta} = \{S | S \in \lambda(\varphi) \bigcup \overline{\lambda(\theta)}\}, l = 1, 2, ..., t.$$

When $\mathbb{S}^{\varphi \wedge \neg \theta} = \emptyset$,

$$D(\theta, \varphi) = \frac{1}{pr} \sum_{i=1}^{r} \sum_{j=1}^{p} D(S_i^{\theta \wedge \neg \varphi}, S_j^{\varphi}) \tag{14}$$

[1]if $dim(\mathbf{x}) > 2$, we define: $\mathbf{x}_1 \geq \mathbf{x}_2$ if $d(\mathbf{x}_1, 0) \geq d(\mathbf{x}_2, 0)$, where $d(\cdot)$ is the Euclidean distance.

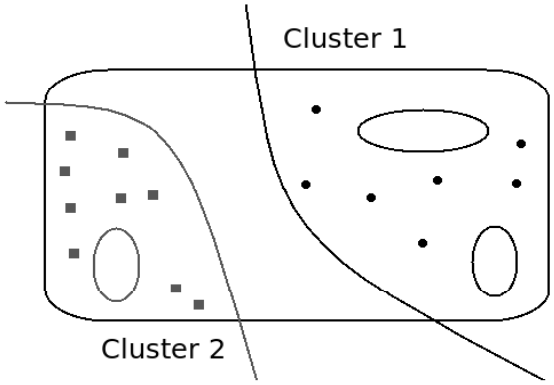

Fig. 2. A schematic illustration of mixed objects clustering. The imprecise concepts are represented by ellipses and data instances are by dots.

When $\mathbb{S}^{\theta\wedge\neg\varphi} = \emptyset$,

$$D(\theta,\varphi) = \frac{1}{qt}\sum_{k=1}^{t}\sum_{l=1}^{q} D(S_k^{\varphi\wedge\neg\theta}, S_l^{\theta}) \tag{15}$$

The above logical expression is one dimensional, which can be used by agents to describe one of the features of the object. If we have a multi-dimensional objects, linguistic rule can be used to described the object such as "x is *big* AND y is *medium* $\wedge$ *large*".

Based on Definition 1, a linguistic rule is a rule that can be represented as a multi-dimensional logical expressions of linguistic labels.

Definition 7: (Multi-dimensional logical expressions of labels) $MLE^{(n)}$ is the set of all multi-dimensional label expressions that can be generated from the logical label expression $LE_j : j = 1, ..., n$ and is defined recursively by

(i) If $\theta \in LE_j$ for $j = 1, ..., n$ then $\theta \in MLE^{(n)}$.

(ii) If $\theta, \varphi \in MLE^{(n)}$ then $\neg\theta, \theta \wedge \varphi, \theta \vee \varphi, \theta \rightarrow \varphi \in MLE^{(n)}$

Similarly we could give the definition of distance between two $MLE^{(n)}$.

Definition 8: (Distance between multi-dimensional logical expressions) Given two n-dimensional logical expressions: Φ, Ψ with

$$\Phi = \theta_{D_1} \wedge \theta_{D_2} \wedge ... \wedge \theta_{D_n}$$

$$\Psi = \varphi_{D_1} \wedge \varphi_{D_2} \wedge ... \wedge \varphi_{D_n}$$

where θ_{D_i}, φ_{D_i} respectively means the logical expressions in dimension D_i. hence, the distance between Φ and Ψ is defined as follows:

$$D(\Phi,\Psi) = \sqrt{\sum_{i=1}^{n} |D(\theta_{D_i}, \varphi_{D_i})|^2} \tag{16}$$

TABLE I
THE DESCRIPTION OF LOGICAL DISTANCE BASED K-MEANS ALGORITHM

Given unlabeled data set DB=$\{\mathbf{obj}_i : i = 1, ..., N\}$ in $\mathbb{R}^d$ and cluster number $K > 0$, counter $p = 0$, $\varepsilon > 0$
Randomly initialize the K centers $\mathbf{c}_1$,..., $\mathbf{c}_K$

While(TRUE): $p++$

Step 1: For each $i \in \{1, ..., N\}$, determine the cluster for all objects in DB: $\mathbf{obj}_i \leftarrow c_j$, if:
$D(\mathbf{obj}_i, \mathbf{c}_j^{(p-1)})$=min$\{D(\mathbf{obj}_i, \mathbf{c}_k^{(p-1)}) : k = 1, ..., K\}$

Step 2: Compute the cluster centers $\mathbf{c}_i^{(p)}$ which satisfy:
$\sum_{\mathbf{obj}_j \in \mathbb{C}_i} D(\mathbf{c}_i^{(p)}, \mathbf{obj}_j)$=min$\{\sum_{\mathbf{obj}_j \in \mathbb{C}_i} D(\mathbf{x}, \mathbf{obj}_j)\}$

Until $|\mathbf{c}^{(p)}$-$\mathbf{c}^{(p-1)}| < \varepsilon$.

IV. CLUSTERING OF MIXED OBJECTS

Based on the above dissimilarity measure, the distance between objects including numerical data, concepts, and linguistic description can be easily measured. Further more, a set of these objects can be grouped into clusters using clustering algorithms.

K-means [8] is one of the simplest unsupervised learning algorithms that solve clustering problem. The procedure follows a simple way to group a given data set to a certain number of clusters. Suppose that we have N sample feature vectors $\langle \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \rangle$ that are all from the same class, and we know that they fall into K ($K < N$) compact clusters, . Let $\mathbf{m}_j$ be the mean of the vectors in cluster j. We can use a minimum-distance to separate them. That is, we can say $\mathbf{x}$ belongs to cluster j if $||\mathbf{x} - \mathbf{m}_j||$ is the minimum of all the K distances. Hence, this algorithm aims at minimizing an objective function J:

$$J = \sum_{j=1}^{K}\sum_{i=1}^{N} ||\mathbf{x}_i^{(j)} - \mathbf{m}_j||^2 \tag{17}$$

Fig. 2 gives a schematic illustration of mixed objects clustering where the ellipses represent some imprecise concepts can could be interpreted into linguistic labels or other granularity. Such mixed objects can be clustered by combining the classical clustering algorithms with our proposed distance measure. Formally, suppose we are given an unlabeled data set $\mathbf{obj}_1, ..., \mathbf{obj}_n$ in which there are numerical data, labels (concepts), and logical expressions (linguistic description). Now we focus on clustering them using K-means algorithm. The algorithm is as same as the classical K-means clustering and the pseudo-code are shown in Table I.

There are two points which are worthy of highlighting. First and the most important, we adopted the new proposed distance metric in last section as the distance measure of these mixed objects. The novel distance definition makes it possible to cluster data and imprecise concepts represented by linguistic labels. Second, in the algorithm implementation, the center of each cluster should be numerical data but not logical expressions.

TABLE II
DISTANCES BETWEEN SETS OF LABELS DEFINED ON H VALUES.

	$\{r\}$	$\{r,o\}$	$\{o\}$	$\{o,y\}$	$\{y\}$
$\{r\}$	0	0.0288	0.0626	0.0950	0.1246
$\{r,o\}$	0.0288	0	0.0354	0.0673	0.0968
$\{o\}$	0.0626	0.0354	0	0.0336	0.0626
$\{o,y\}$	0.0950	0.0673	0.0336	0	0.0304
$\{y\}$	0.1246	0.0968	0.0626	0.0304	0

TABLE III
DISTANCES BETWEEN EACH TWO LOGICAL EXPRESSIONS θ, φ AND γ DEFINED ON H VALUES.

	$\theta = \neg orange$	$\varphi = red \vee orange$	$\gamma = red$
$\theta = \neg orange$	0	0.1413	0.1735
$\varphi = red \vee orange$	0.1413	0	0.0650
$\gamma = red$	0.1735	0.0650	0

V. EXPERIMENTAL STUDIES

In this section we did two experiments to verify the new proposed measure and the clustering problem. The first experiment is to illustrate the properties of the proposed logical distance measure. The second experiments shows image scene clustering based on a benchmark problem.

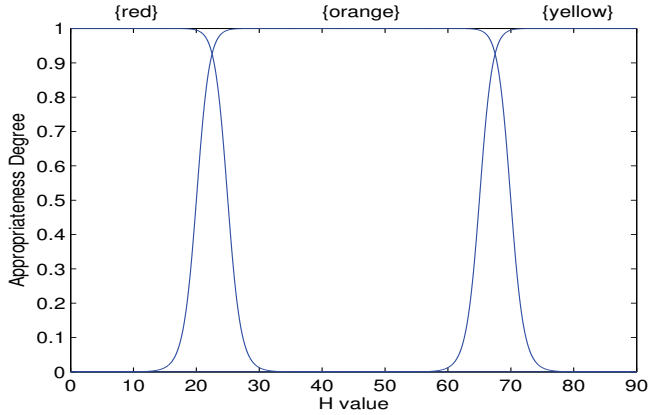

Fig. 3. Three linguistic labels: red (r), orange (o) and yellow (y). They are defined by bell-shape fuzzy sets on the Hue (H) values.

A. Logical Distance

Color of an object is a vague concept and largely depends on the observer's subjective belief. Human eye can distinguish about 10 million different colors but we only use very limited number of words to describe them. There is no general principle for human beings to decide using the most appropriate word such as yellow or red to describe the color of an object. However, people know the difference among these colors, which is difficult to be defined with some specific numerical value. We often make classifications based on these vague and subjective differences among objects. We can use the method proposed in this research to quantize these differences in order to let the agents have the ability to distinguish different objects.

HSV color space can be well visualized by the conical representation model accords with human beings' visual features. Fig. 3 shows three linguistic labels defined on the Hue (H)

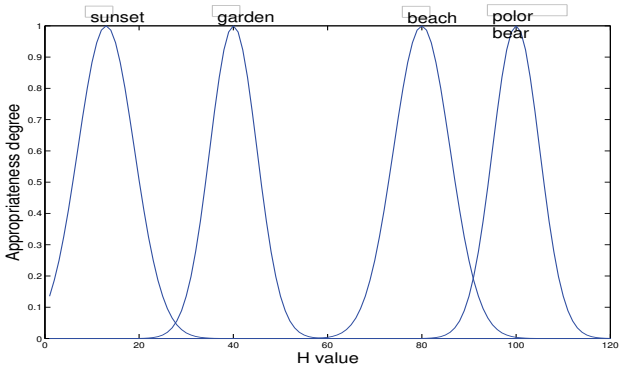

Fig. 4. Four image labels are predefined on the HSV space. This is the mapping of the labels on Hue (H) axis.

value. Based on Definition 3, we can infer that the focal set is:

$$\mathbb{F} = \{\{r\}, \{r,o\}, \{o\}, \{o,y\}, \{y\}\}$$

The distances between these label sets based on Definition 5 is calculated and shown into Table II, from which we can clearly see the symmetric property of this distance measure.

Given the same example above, if there is imprecise concept of the color is not orange, or formally denoted by θ: $\theta = \neg orange$. There are also two other descriptions about the colors φ and γ:

$$\varphi = red \vee orange,\ \gamma = red$$

According to Definition 2, the possible label sets of the given logical expressions θ, γ and φ are calculated as follows:

$$\lambda(\neg o) = \{\{r\}, \{y\}\},\ \lambda(r) = \{\{r\}, \{r,o\}\}$$
$$\lambda(o) = \{\{r,o\}, \{o\}\,\{o,y\}\}$$

so that

$$\lambda(\theta) = \{\{r\}, \{y\}\},\ \lambda(\gamma) = \{\{r\}, \{r,o\}\}$$
$$\lambda(\varphi) = \{\{r\}, \{r,o\}, \{o\}\,\{o,y\}\}$$

Then according to Definition 6 and Table II, we could calculate all the distances and fill in Table III. It is illustrated that distances between LEs are symmetric and could reasonably reflect the logical divergences between LEs.

B. Images and Labels Clustering

To show how to use our proposed approach to cluster data and imprecise concept, we built a toy application which can cluster images and image labels which are defined based on image global color feature. One hundred images are evenly picked up from four categories of the Corel image data set [2][7]. And each category has 25 images. We first resize these images into 192×128. We then artificially designed four linguistic labels on color: "$sunset$", "$beach$", "$garden$", and "$polar\ bear$" based on the image content. These labels can be considered as the linguistic descriptors of images which convey the semantic of images. Since the main goal of this

Fig. 5. Sample results of images and linguistic labels clustering. Linguistic labels on color: *sunset, garden, beach* and *polar bear* are areas of color in HSV space. For each cluster, four images and one label (the colored shape on the top of stack of images) are shown.

TABLE IV
PERFORMANCE OF THE IMAGE CLUSTERING.

	sunset	*beach*	*garden*	*polar bear*
Performance	72%	60%	64%	96%

work is to present a new approach to cluster data and vague concepts, we did not use complicated image features like SIFT and LBP.

The only feature we used in this experiment is the global average of HSV color. Fig. 4 shows the mappings of these predefined labels on the Hue (H) axis. The original labels are 3-dimensional granularity in the HSV space. Thus after artificially designing the image labels and extracting image color feature, the data set need to be clustered is with one hundred images and the above four image labels which are represented by granularity on HSV space. The experimental results are shown in Table IV and the code and image set are also open to public at the web address of [18] and [19], respectively. In the future work, such linguistic labels can be used as constraints guide clustering process by using high-level knowledge. It allows us to develop a new human-machine interface by using linguistic labels.

In addition although low-level image features, such as color and texture, are useful information to describe images in this experiment. In practice the features extraction are very important and the clustering algorithms are always only one step of the procedure. High-level image features can convey much more accurate information of the content of the images. In real-world application of image clustering, SIFT, LBP or other complex image features are better choice.

VI. CONCLUSION

In this paper we proposed a novel clustering algorithm by employing a new distance measure based on label semantics. The new algorithm can be used to cluster data and imprecise concepts. The distance defined in linguistic label space differs from the other measures by focusing on the difference of logical meanings the objects convey. It has the ability to measure the linguistic divergence between numerical data and concepts which are presented in the form of linguistic labels. Experimental studies on a toy image clustering problem showed that our approach is an effective to group data and linguistic labels reasonably.

Because the new proposed measure is calculated by considering the accumulated dissimilarities alone the trials connecting to imprecise concepts. It is extendable to measuring distance between any granularities. In the future work, we hope to apply this clustering algorithms to other application areas beside image classification. We will also investigate the advantages and disadvantages of this measure studying other concepts represented by other granularity.

ACKNOWLEDGEMENTS

This work is partially funded by the NCET Program of MOE, China and the SRF for ROCS. The second author also thanks the China Scholar Council for visiting fellowship (No. 2010307502).

REFERENCES

[1] M.M.S. Beg, M. Thint and Z. Qin, PNL-enhanced restricted domain question answering system, *the Proceedings of IEEE-FUZZ*, (2007) pp. 1277-1283.
[2] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29 (3) (2006), pp. 394-410.
[3] L.K. Hyung, Y.S. Song, K.M. Lee, Similarity measure between fuzzy sets and between elements, *Fuzzy Sets and System* Vol. 62 (1994), pp. 291-293.
[4] J. Lawry, *Modelling and Reasoning with Vague Concepts* (2006), Springer.
[5] J. Lawry, A Framework for Linguistic Modelling, *Artificial Intelligence*, Vol. 155 (2004), pp. 1-39
[6] D.-F. Li, Some measures of dissimilarity in intuitionistic fuzzy structures, *Journal of Computer and System Sciences* Vol. 8 (2004), pp. 115-122.
[7] V. Lavrenko, R. Manmatha, and J. Jeon, A model for learning the semantics of pictures, *Proceedings of NIPS*, (2004).
[8] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press (1967), pp. 281-297.
[9] Z. Qin, J. Lawry, Decision tree learning with fuzzy labels, *Inform. Sci.* 172(1-2) (2005), pp. 91-129.
[10] Z. Qin, M. Thint and M.M.S. Beg Deduction engine designs for PNL-based question answering systems, *Foundations of Fuzzy Logic and Soft Computing*, LNAI 4529 (2007) pp. 253-262.
[11] Z. Qin, J. Lawry, LFOIL:Linguistic rule induction in the label semantics framework, *Fuzzy sets and systems* Vol. 159 (2008), pp. 435-448.
[12] Z. Qin, J. Lawry, Knowledge discovery in a framework for modelling with words. *Soft Computing for Knowledge Discovery and Data Mining* (2008): pp. 241-276.
[13] T. Wan and Z. Qin, A new technique for summarizing video sequences through histogram evolution, *Proceedings of Int. Conf. on Signal Processing and Communications (SPCOM)* (2010), pp. 1-5.
[14] L.A. Zadeh, Fuzzy Logic=Computing with Words, *IEEE Transaction on Fuzzy Systems* Vol. 4(1996), pp. 103-111.
[15] L.A. Zadeh, The Concept of Linguistic Variable and its Application to Approximate Reasoning Part 2, *Information Science* Vol. 8 (1975), pp. 301-357.
[16] W. Zhang, and Z. Qin, Dissimilarity measure of logical expressions, *Proceedings of Int. Conf. of Machine Learning and Cybernetics (ICMLC)* (2010), pp. 199-203.
[17] http://en.wikipedia.org/wiki/Unsupervised_learning
[18] icmll.buaa.edu.cn/projects/FUZZ-IEEE2011/code.rar
[19] icmll.buaa.edu.cn/projects/FUZZ-IEEE2011/image.rar