# MembershipMap: Data Transformation Based on Membership Aggregation

Hichem Frigui
University of Memphis, hfrigui@memphis.edu

## Abstract

*We propose a new data-driven transformation that facilitates many data mining, interpretation, and analysis tasks. Our approach, called MembershipMap, strives to extract the underlying sub-concepts of each raw attribute, and uses the orthogonal union of these sub-concepts to define a new space. The sub-concept soft labels of each point in the original space determine the position of that point in the new space. Since sub-concept labels are prone to uncertainty inherent in the original data and in the initial extraction process, a combination of labeling schemes that are based on different measures of uncertainty will be presented. In particular, we introduce the CrispMap, SoftMap, and PossibilisticMap. We show that the MembershipMap can be used as a flexible pre-processing tool to support such tasks as: sampling, data cleaning, and outlier detection.*

## 1. Introduction

Knowledge Discovery in Databases (KDD) aims at discovering interesting and useful knowledge from large amounts of data. The KDD process consists of three main phases[3]: data preparation and preprocessing, data mining, and pattern evaluation and presentation. The preprocessing phase, is an essential but compared to the other phases, largely ignored subject. Traditionally, it has taken a backseat to the data mining algorithms. However, without adequate preparation of the data, the outcome of any data mining algorithm can be disappointing, hence the saying "garbage in garbage out".

In this paper, we propose a data transformation that facilitates many KDD tasks. Our approach, called MembershipMap, takes into account the intricate nature of each attribute's distribution *independently* and aggregates the soft cluster labels to create a new membership space. The MembershipMap has several advantages including: **(i)** The mapping can be exploited in any of the three complementary spaces: crisp, soft, or possibilistic; **(ii)** data can be easily identified as noise, boundary, or seeds.

Even though the attributes are treated independently, they do get combined at a later stage. Hence, information such as correlations is not lost in the transformed space. In fact, treating the attributes independently in the initial phase is analogous to mean-variance normalization, and if the data has a simple unimodal distribution along each dimension, the two approaches are equivalent. However, since in general, multiple clusters are sought in each dimension, the MembershipMap is essentially a generalization of the simple normalization that performs local scaling that is optimal for each sub-group or cluster.

## 2. Data Partitioning and Labeling

Clustering aims at classifying a set of $N$ unlabeled data points $\mathcal{X}=\{\mathbf{x}_j|j=1,\ldots,N\}$ into $C$ clusters $G_1,\cdots,G_C$, and assigning a label to each point to represent information about its belongingness. In general, there are four types of cluster labels depending on the uncertainty framework: **crisp**, **fuzzy**, **probabilistic**, and **possibilistic**. Each labeling paradigm uses a different uncertainty model, and thus, have a different interpretation. In crisp labeling, each $\mathbf{x}_j$ is assigned a binary membership value,$(u_c)_{ij}$, such that:

$$(u_c)_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in G_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Fuzzy and probabilistic labeling, allow for partial or gradual membership values. This labeling offers a richer representation of belongingness and can handle uncertain cases. Fuzzy and probabilistic memberships can have a different interpretation at high level reasoning. However, for the purpose of assigning labels, they can be treated the same way. We will refer to both of them as *(constrained) Soft* labels. Soft labels, $(u_f)$, satisfy the constraints:

$$0 \leq (u_f)_{ij} \leq 1, and \sum_{i=1}^{C}(u_f)_{ij} = 1,\ \forall j = 1\cdots N. \quad (2)$$

Many fuzzy clustering algorithms assign this type of labels. For instance, the Fuzzy C-Means (FCM) [1] uses:

$$(u_f)_{ij} = \frac{(1/d_{ij})^{\frac{2}{m-1}}}{\sum_{k=1}^{C}(1/d_{kj})^{\frac{2}{m-1}}}, \quad (3)$$

where $d_{ij}$ is the distance between cluster $i$ and $\mathbf{x}_j$, and $m\in(1,\infty)$ controls the degree of fuzziness.

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

In possibilistic labeling the constraint that the memberships must sum to one is relaxed. This framework assigns *"typicality"* values, $(u_p)$, that do not consider the *relative* position of the point to all clusters. Many robust clustering algorithms [4, 6] use this type of labeling. For instance, the Possibilistic C-Means [6] uses

$$(u_p)_{ij} = \frac{1}{1 + (d_{ij}/\eta_i)^{\frac{2}{m-1}}}, \tag{4}$$

where $\eta_i$ is a cluster-dependent resolution [6]. Robust statistical estimators, such M- and W-estimators [5] use this type of memberships to reduce the effect of noise.

## 3. MembershipMap Generation

We propose a data transformation that maps the original feature space into a membership unit hypercube with richer information content. This transformation starts by clustering each attribute independently. Each cluster, thus obtained, can be considered to form a "subspace" that reflects a "more specific" concept along that dimension. The orthogonal union of all the subspaces compose the transformed space, which we refer to as *MembershipMap*. The class labels (crisp, soft, or possibilistic) determine the position of data points in the new space.

The MembershipMap can be mined just like the original data space, for clusters, classification models, association rules, etc. This task is expected to be easier in the membership space because all features are confined to the interval [0,1], and have special meaning within this interval. Moreover, since different areas of the transformed spaces correspond to different concepts, the MembershipMap can be "explored" in a pre-data mining phase to uncover useful underlying information. In this paper, we focus our attention in exploring the soft and possibilistic Membership spaces, which we will refer to as SoftMap and PossibilisticMap respectively. We outline how these maps could be used to uncover seed points; noise and outliers; and boundary points.

### 3.1. Identifying Seed Points

In applications involving huge amounts of data,"*Seed points*" identification may be needed to reduce the complexity of data-driven algorithms. This process can be difficult, if not impossible, for data with uneven, noisy, heterogeneous distributions. In addition to sampling, seed points can offer excellent initialization for techniques that are sensitive to the initial parameters such as clustering.

Using the PossibilisticMap, seed points can be identified as points with high typicality, where typicality is defined as

$$T_{\mathbf{x}_j} = \min_{k=1,\cdots,n} \Big\{ \underbrace{\max_{l=1,\cdots,C_k} \{u_{lj}^{(k)}\}}_{\text{Best Possibilistic label in Dim k}} \Big\}. \tag{5}$$

### 3.2. Identifying Noise Points and Outliers

Noise and outlier detection is a challenging problem. In fact, when data has a large, unknown proportion of outliers, most learning algorithms break down and cannot yield any useful solution. Using the PossibilisticMap, Noise and outliers can be identified as those points located near the origin. In other words, if $\mathbf{x}_j^P$ is the map of $\mathbf{x}_j$ in the PossibilisticMap, then $\mathbf{x}_j$ is a noise point if $\|\mathbf{x}_j^P\|$ is small.

### 3.3. Identifying Boundary Points

The process of boundary points identification is paramount in many classification techniques. It can be used to reduce the training data to (or emphasize) a set of boundary points to achieve higher focus in learning class boundaries. Using the SoftMap, boundary points can be identified as points that do not have a strong commitment to any cluster of a given attribute. In other words, they are points that belong to multiple units and have low purity values, where the purity of $\mathbf{x}_j$, $P_{\mathbf{x}_j}$, is defined as

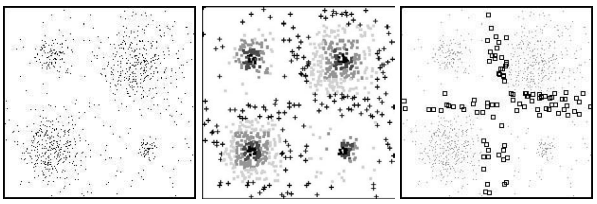
$$P_{\mathbf{x}_j} = \min_{k=1,\cdots,n} \Big\{ \underbrace{\max_{l=1,\cdots,C_k} \{u_{lj}^{(k)}\}}_{\text{Best Soft label in Dim k}} \Big\}. \tag{6}$$

Fig.1 displays a simple data set with 4 clusters. First, the projection of the data along each dimension is partitioned into 2 clusters using the FCM [1]. Next, possibilistic labels were assigned to each $\mathbf{x}_j$ in all 4 clusters using (4), and each $\mathbf{x}_j$ is mapped to $\mathbf{x}_j^P$. In Fig. 1(b), we have quantized $T_{\mathbf{x}_j}$ into 4 levels: $T_{\mathbf{x}_j} \geq 0.95$ (black squares); $0.80 \leq T_{\mathbf{x}_j} < 0.95$; $0.50 \leq T_{\mathbf{x}_j} < 0.80$; and $0.25 \leq T_{\mathbf{x}_j} < 0.50$ (squares with shade that gets lighter). As can be seen, points located at the core of each cluster have the highest degree of typicality and could be treated as "seeds". The degree of typicality decreases as we move away from the clusters. Points near the origin, i.e., $\|\mathbf{x}_j^P\|$ is small ($<0.25$), are identified and displayed using the "+" symbol in Fig.1(b). These are generally noise points. To identify boundary points, soft labels were assigned to each $\mathbf{x}_j$ in all 4 clusters using (3), and each $\mathbf{x}_j$ is mapped to $\mathbf{x}_j^S$. Fig. 1(c) displays the points $\{\mathbf{x}_j \mid P_{\mathbf{x}_j} < 0.75\}$ as small squares. These are mainly the cluster boundaries.

## 4. Experimental Results

### 4.1. Identifying Regions of Interest

We illustrate the ability of the proposed transformation to identify regions of interest using the Iris data set. The first step in the MembershipMap transformation consists of a quick and rough segmentation of each feature. This is a relatively easy task that can rely on histogram thresholding or clustering 1-D data. The results reported in this paper were obtained using the FCM with the number of clusters



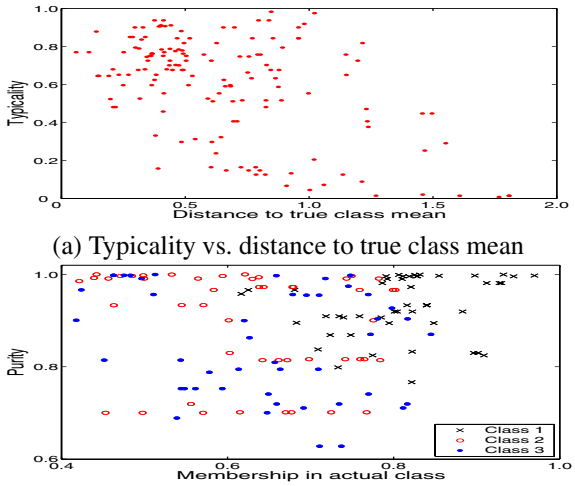
(a) (b) (c)

**Figure 1.** **Exploring the MembershipMaps. (a) Data set. Points with (b) different typicality, (c) low purity.**

fixed to 3 for each dimension. Next, fuzzy and possibilistic labels were assigned to each sample using equations (3) and (4). Finally, the fuzzy (possibilistic) labels were aggregated to create the FuzzyMap (PossibilisticMap). Each map has 12 dimensions.

To identify seed points, noise and outliers, we compute the typicality of each sample. Points with high typicality would correspond to seed points, while points with low typicality would correspond to noise points. Since the Iris data has 4 dimensions, the identified points could not be visualized as in Fig. 1. Instead, we rely on the ground truth to validate the results. We use the class labels and compute the centroid of each class. Then, we compute the distance from each point to the centroid of its class. Theoretically, points with small distances would correspond to seeds, and points with large distances correspond to noise. Fig. 2(a) displays a scatter plot of the sample typicality versus their distance to the true class centroid. The distribution clearly suggests a negative correlation: Typicality, which was extracted only from the PossibilisticMap *without any knowledge* about the true class labels, is higher for those samples located near the true class centroid.

To validate the degree of purity, we use the ground truth and compute the membership of each sample using eq.(3), where the distances are computed with respect to the classes' centroids. Fig. 2(b) displays a scatter plot of the sample purity versus their membership in the true class. The distribution shows a positive correlation indicating that the purity computed using the SoftMap without any knowledge about the true class labels can estimate the degree of sharing among the multiple classes, i.e, boundary points. Moreover, we notice that all samples from class 1 have high purity, and the few samples with low purity belong to either class 2 or 3. This information is consistent with the known distribution of the Iris data where class 1 is well-separated while classes 2 and 3 overlap. Fig. 3 displays typical segmentation results. The first raw displays the original images to be segmented. Each image is of size 128 by 192. Thus, the segmentation process involves clustering 24,576 points in a 6-D feature space. In all examples shown in this paper, we use the FCM clustering algoritn, fix the number of




(a) Typicality vs. distance to true class mean

(b) Purity vs. membership to true class

**Figure 2.** **Validating identified regions of interest**

clusters to 5, and fix m to 2.0. Fig. 3(b) displays the segmentation results when all pixels are clustered.

### 4.2. Application: Color Image Segmentation

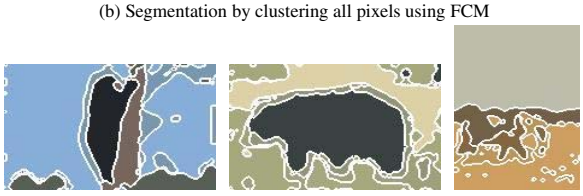
The proposed transformation and pre-processing techniques are applied to the problem of color image segmentation. Each pixel in the image is mapped to a 6-D feature vector [2] consisting of 3 color and 3 texture features.

To extract regions of interest for the purpose of data reduction and cleaning, we first construct the fuzzy and possibilistic maps for the feature vectors of each image as previously described for the Iris data. The fuzzy and possibilistic maps of each image have 18 dimensions. For each image, we use the *PossibilisticMap* to identify seed and noise points. Seed points are selected as the top 5% of the points having the highest typicality values (using eq. (5)), while noise points are the bottom 5% of the points having the lowest typicality values. To identify boundary points, we use the *FuzzyMap* and select 5% of the points having the lowest purity values (using eq. (6)). We should note here that our choice of 5% is arbitrary. We could have used a lower or higher percentage, or we could have thresholded the typicality and purity values. The identified points of interest are shown in Fig. 4. As can be seen, seed points correspond to typical points from the different regions in the image. Noise points, on the other hand, correspond to edges (due to the smoothing during feature extraction) and small areas with different color and texture features (e.g. small tree branches). Boundary points are harder to interpret in the Figure. However, many of them correspond to points that separate different regions.

To illustrate the sampling feature of the *PossibilisticMap*, we cluster only the seed points (1,200 points, compared to

(a) Original Images

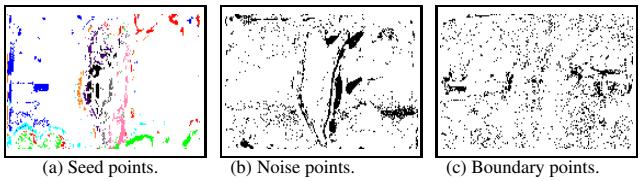
(b) Segmentation by clustering all pixels using FCM



(c) Segmentation by clustering seed points only (5%) using FCM

(d) Segmentation by clustering pre-processed data using FCM

**Figure 3.** **Segmentation Results**



(a) Seed points. (b) Noise points. (c) Boundary points.

**Figure 4.** **Identified points of interest for the first image in Fig. 3(a)**

clustering 24,576 points). Then, all image pixels are assigned to the closest seed cluster. Fig. 3(b) displays these segmentation results. As can be seen, the results are comparable, even though these segmented images were obtained in a fraction of the time required to cluster all pixels. In fact, for the first two images, segmentation with seeds only yields slightly better results. The improved performance may be due to the fact that seed points constitute a much cleaner data with compact and well-separated clusters, which makes the FCM less susceptible to local minima. The drawback of clustering seedpoints only is that there is no guarantee that enough seed points would be identified from each region to form a valid cluster. This might explain the reason that the back of the elephant was merged with part of the background in Fig. 3(c).

To illustrate the data cleaning feature of the Membership Maps, we exclude noise and boundary points from the clustering process, and use seed points to initialize the prototypes. The segmentation results are shown in Fig. 3(d). As can be seen, data cleaning improves the segmentation results. For instance, the back of the elephant is no longer confused with the background.

## 5. Conclusions

We have presented a new mapping that facilitates many data mining tasks. Our approach strives to extract the underlying sub-concepts of each attribute, and uses their orthogonal union to define a new space. The sub-concepts of each attribute can be easily identified using simple 1-D clustering or histogram thresholding. Moreover, since fuzzy and possibilistic labeling can tolerate uncertainties and vagueness, there is no need for accurate sub-concept extraction. In addition to improving the performance of clustering and classification algorithms by taking advantage of the richer information content, the MembershipMaps could be used to formulate simple queries to extract special regions of interest, such as noise, outliers, boundary, and seed points.

There is a natural trade-off between dimensionality and information gain. Increased dimensionality of the MembershipMap can be offset by the quality of hidden knowledge that is uncovered, and can even be avoided by going back to the original feature space after data reduction and cleaning as illustrated in the image segmentation application. Thus, benefiting both from lower dimensionality and lower data cardinality.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0133415.

## References

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[2] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer, 1999.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

[4] H. Frigui and R. Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *IEEE Trans. Patt. Analysis Mach. Intell.*, 21(5):450–465, 1999.

[5] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics the Approach Based on Influence Functions*. John Wiley & Sons, New York, 1986.

[6] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Systems*, 1(2):98–110, 1993.