Pattern Recognition Letters 56 (2015) 1–6

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# A new perspective of modified partition coefficient ☆

Abdul Suleman*

*Instituto Universitário de Lisboa (ISCTE - IUL), BRU, Av. Forças Armadas, Lisbon, Portugal*

ARTICLE INFO

*Article history:*
Received 21 July 2014
Available online 2 February 2015

*Keywords:*
Fuzzy clustering
Fuzzy partition
Partition coefficient
Divergence measures

ABSTRACT

We begin by showing that the modified partition coefficient (MPC) is an average Euclidean distance between membership degrees and the centre of the fuzzy *c*-partition. Subsequently, we construct alternative MPCs using several other measures of dissimilarity and examine how differently they perform when compared with the original proposal. Empirical evidence shows that the MPC based on a Chernoff's measure of divergence is more robust to the initial conditions of the fuzzy *c*-means algorithm.



## 1. Introduction

Consider $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ a set $N$ data points represented by $n$-dimensional feature vectors, i.e. $\mathbf{x}_k \in \mathbb{R}^n,\ 1 \le k \le N$. We represent this set as an $n \times N$ matrix $\mathbf{X}$. Each column of this matrix represents a data point, and the rows of $\mathbf{X}$ account for the $n$ features. By assumption, $\mathbf{X}$ can be decomposed into unknown $c \in (1, N)$ fuzzy clusters forming a fuzzy $c$-partition. A fuzzy $c$-partition can be appropriately represented as a $c \times N$ matrix denoted by $\mathbf{U} = [\mu_{ik}]$, and referred to as partition matrix. We also assume this matrix is given as an output of the algorithm used to decompose $\mathbf{X}$. Its generic element $\mu_{ik}$ is the membership degree of $\mathbf{x}_k$ in fuzzy cluster $i$. The columns of $\mathbf{U}$, i.e. the vectors

$$\mu_k = (\mu_{1k},\ \mu_{2k},\ \ldots,\ \mu_{ck}) \quad (1)$$

for $1 \le k \le N$, belong to the unit simplex $\mathcal{S}_c$,

$$\mathcal{S}_c = \left\{ (g_1, \ldots, g_c) : g_i \ge 0, \sum_{i=1}^{c} g_i = 1 \right\} \quad (2)$$

This convex set is a geometrical counterpart of the fuzzy $c$-partition. The cluster full members or prototypes are represented by the canonical basis vectors of $\mathbb{R}^c$, and consequently locate at the vertices or extreme points of $\mathcal{S}_c$. We call

$$\mathcal{C} = \left( \frac{1}{c}, \frac{1}{c}, \cdots, \frac{1}{c} \right) \quad (3)$$

the centre of the fuzzy $c$-partition or, equivalently, of $\mathcal{S}_c$; it corresponds to the fuzziest point [10].

Now suppose we have a collection of fuzzy $c$-partitions of $\mathbf{X}$, e.g. $c = 2, 3, \ldots$, and aim to evaluate how well each partition fits the data; in other words, we want to know how to select the best estimate among candidates that partition $\mathbf{X}$. Bezdek [5] proposes a scalar measure or cluster validity index, called partition coefficient (PC), given by

$$\mathcal{V}_{\text{PC}}(c) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{c} \mu_{ik}^2 \quad (4)$$

to assess competing fuzzy partitions of $\mathbf{X}$. It can be shown that $\mathcal{V}_{\text{PC}} \in \left[\frac{1}{c}, 1\right]$. The lower extreme is attained if $\mathbf{U} = \left[\frac{1}{c}\right]$, while the value $\mathcal{V}_{\text{PC}} = 1$ holds if the partition is hard, i.e. when all data points are mapped into the vertices of the unit simplex $\mathcal{S}_c$. Pal and Bezdek [18] interpret this verbally, noting that $\mathcal{V}_{\text{PC}}$ measures how far $\mathbf{U}$ is from being a crisp partition matrix. The higher the values of $\mathcal{V}_{\text{PC}}$, the harder $\mathbf{U}$ is. Despite its appealing conceptual construction and simplicity, the index $\mathcal{V}_{\text{PC}}$ has the drawback of tending to increase monotonically with the number of clusters $c$.

Dave [9] proposed a modification of (4), by means of a linear transformation, to eliminate that dependency of $\mathcal{V}_{\text{PC}}$ on $c$. He refers to the new measure thus obtained as modified partition coefficient (MPC), which is expressed as

$$\mathcal{V}_{\text{MPC}}(c) = \frac{c}{c-1} \mathcal{V}_{\text{PC}} - \frac{1}{c-1} \quad (5)$$

and is similar to the clustering performance measure of [3]. The range of $\mathcal{V}_{\text{MPC}}$ is the unit interval $[0, 1]$, where $\mathcal{V}_{\text{MPC}} = 0$ corresponds to maximum fuzziness and $\mathcal{V}_{\text{MPC}} = 1$ to a hard partition. This feature makes $\mathcal{V}_{\text{MPC}}$ additionally more attractive than the Bezdek's PC to compare different cluster solutions for the data matrix $\mathbf{X}$, since it does not depend on the number of clusters as does the range of $\mathcal{V}_{\text{PC}}$ (4). In general, the optimal number of clusters is found by solving $\max_c \mathcal{V}_{\text{MPC}}(c)$.

During our research we found that the MPC (5) can be obtained in a more appealing way, which allows this performance measure to be looked on as a distance of cluster memberships to the fuzziest point, as follows.

☆ This paper has been recommended for acceptance by Andrea Torsello.

* Tel.: +351917244744.
*E-mail address:* abdul.suleman@iscte.pt

http://dx.doi.org/10.1016/j.patrec.2015.01.008
0167-8655/

**Proposition 1.** *The modified partition coefficient (5) is an average of a normalised squared Euclidean distance of membership degree vectors (1) to the centre of the fuzzy c-partition (3).*

**Proof.** Note that the squared Euclidean distance to the centre of the unit simplex $\mathcal{S}_c$ (2) of any vertex is equal to $1 - \frac{1}{c} = \frac{c-1}{c}$. Thus, the squared Euclidean distance between any membership degree vector and the centre, multiplied by $\frac{c}{c-1}$, gives a number in the range [0, 1], i.e.

$$\frac{c}{c-1}\sum_{i=1}^{c}\left(\mu_{ik}-\frac{1}{c}\right)^2=\frac{c}{c-1}\sum_{i=1}^{c}\mu_{ik}^2-\frac{1}{c-1}$$

Summing this quantity for all data points, and dividing by $N$, gives the MPC index (5) proposed by Dave [9]. □

In other words, the fuzzy $c$-partition that best represents $\mathbf{X}$ is the one that, in general, leads the data points the furthest from the centre of $\mathcal{S}_c$. This new perspective of MPC provided by Proposition 1 raises the following question: is there any dissimilarity or divergence measure we can use instead of the Euclidean distance to construct alternative MPCs that better addresses the cluster validation problem? Our study attempts to answer this question. In what follows, we revisit divergence measures in Section 2 and associate each one with a cluster validity index; in Section 3 we carry out an empirical study using first synthetic data and then datasets from UCI Machine Learning Repository [2]; and finally Section 4 concludes.

## 2. Measures of divergence

### 2.1. An overview

The similar nature of membership degrees and discrete probabilities makes it possible to use the theoretical measures of divergence developed to compare probability distributions in the present case. The problem can also be regarded as one that compares two multinomial populations [7], where each data point with the associated membership degree vector acts as *a population*, and is compared to the reference population, i.e. the centre of the fuzzy $c$-partition, $\mathcal{C}$ (3). We should then select the partition that generally diverges most from the reference. We stress that we are deliberately using the term *divergence* instead of *distance*, because not all measures of dissimilarity considered here satisfy all the conditions of a distance function. While the latter can always be considered a divergence measure, the converse is not necessarily true [21]. In what follows, we present an overview of a number of divergence measures for two generic multinomial probabilities; then we construct new indices as realisations of MPC for each of such measures. The papers [1,4,21] are the main sources of the first part.

Most coefficients of divergence between probability distributions considered in this study are special cases of Ali-Silvey class of information theoretic measures. They are derived from the general formula [1]

$$d(\mathbf{f}_1,\mathbf{f}_2)=\varphi\left\{\mathbb{E}_1\left[\Lambda\left(\frac{\mathbf{f}_2}{\mathbf{f}_1}\right)\right]\right\} \tag{6}$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are two continuous probability density functions, $\Lambda(.)$ is a continuous real convex function in $\mathbb{R}^+$, $\varphi(.)$ an increasing real function on $\mathbb{R}$, and $\mathbb{E}_1$ denotes the expectation with respect to $\mathbf{f}_1$. The formula (6) gives an account of how $\mathbf{f}_2$ diverges from $\mathbf{f}_1$. In this context, we will however consider the discrete versions of divergence measures, and this corresponds to replacing the symbol of integral by that of summation appropriately. So we consider two generic multinomial populations with the probability vectors $\mathbf{p}_1 = (p_{11}, p_{21}, \ldots, p_{c1})$ and $\mathbf{p}_2 = (p_{12}, p_{22}, \ldots, p_{c2})$, and present several ways to calculate $d(\mathbf{p}_1,\mathbf{p}_2)$. Clearly, the parameter space of these vectors is the unit simplex $\mathcal{S}_c$ (2), i.e. $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{S}_c$.

### 2.2. Examples of divergence measures

With the exception of the arccosin measure [7] below, the coefficients of divergence we present herein result from specifying the functions $\varphi$ and $\Lambda$ in (6). Our list is by no means exhaustive. Various available measures of divergence (e.g. [15]) are members of this generic class, as shown below. The usual nomenclature for the continuous distributions is maintained for the case of discrete probabilities. We notice that whenever the logarithm function is involved, it should be intended as natural logarithm; we use the result $\lim_{a\to 0} a\log(a) = 0$ to set the product $a\log(a)$ to zero, when $a = 0$.

(a) Kullback and Leibler [15] measure of discriminatory information: $\Lambda(a) = -\log(a)$; $\varphi(a) = a$;

$$d(\mathbf{p}_1,\mathbf{p}_2)=\sum_{i=1}^{c}p_{i1}\times\log\left(\frac{p_{i1}}{p_{i2}}\right) \tag{7}$$

This is the most popular measure of divergence between two probability distributions. We note that the coordinates of $\mathbf{p}_2$ must be positive.

(b) Chernoff [8] measure of discriminatory information: $\Lambda(a) = -a^{1-r}$; $\varphi(a) = -\log(-a)$, where $r \in (0, 1)$ is an exponent parameter;

$$d(\mathbf{p}_1,\mathbf{p}_2)=-\log\left(\sum_{i=1}^{c}p_{i1}^{r}\times p_{i2}^{1-r}\right) \tag{8}$$

This measure is very flexible as the value of the exponent parameter $r$ can be adjusted for particular needs. Dividing (8) by $1 - r$ leads to Rényi's [19] divergence[1] of order $r$.

(c) Kolmogorov's variational distance: $\Lambda(a) = |1 - a|$; $\varphi(a) = \frac{1}{2}a$;

$$d(\mathbf{p}_1,\mathbf{p}_2)=\frac{1}{2}\sum_{i=1}^{c}|p_{i1}-p_{i2}|$$

(d) Hellinger distance: $\Lambda(a) = (\sqrt{a} - 1)^2$; $\varphi(a) = \frac{1}{2}a$;

$$d(\mathbf{p}_1,\mathbf{p}_2)=\frac{1}{2}\sum_{i=1}^{c}\left(\sqrt{p_{i1}}-\sqrt{p_{i2}}\right)^2$$

This is also referred to as squared Hellinger distance.

(e) Bhattacharyya's measures of divergence: indeed we can consider two different measures. One is referred to as Bhattacharyya logarithm and the other can be called Bhattacharyya arccosin [16]. The former is a special case of Chernoff's measure (8) with $r = 0.5$ and is treated as such. The arccosin measure proposed in [7] does not belong to Ali-Silvey class referred to above (6), and is given by

$$d(\mathbf{p}_1,\mathbf{p}_2)=\arccos\left(\sum_{i=1}^{c}\left(\sqrt{p_{i1}\times p_{i2}}\right)\right) \tag{9}$$

The author calls (9) the angle of divergence between two populations, as represented by $\mathbf{p}_1$ and $\mathbf{p}_2$.

### 2.3. Alternative MPCs

We have learned from the Proposition 1 that the construction of MPC [9] requires the calculation of the divergence of any extreme point of the unit simplex $\mathcal{S}_c$ to the centre $\mathcal{C}$ (3) of this convex set. The resultant quantity is then used as the normalisation factor for the divergence between the centre and membership degree vector $\mu_k$. Summing this measure for all data points, and dividing by $N$, gives an alternative MPC. Herein we exemplify how it can be obtained for the case of Chernoff's measure. The procedure is exactly the same

[1] In the original work the author uses the logarithm of base 2.