HS-measure : a hybrid clustering validity measure to interpret road traffic data

Yosr Naïja
LIP2, Faculty of Science of Tunis,
Campus Universitaire
2092 El-Manar Tunis, Tunisia
yosr.naija@gmail.com

Kaouther Blibech
LIP2, Faculty of Science of Tunis,
Campus Universitaire
2092 El-Manar Tunis, Tunisia
kaouther.blibech@gmail.com

ABSTRACT

Clustering validity measures aim to evaluate the goodness of clustering results in order to find the best partition. Results are obtained by varying the input parameters values. However, sometimes, the values generated by these measures are very close and the choice of the optimal value associated to the best partition may be meaningless. In this paper, we propose a new concept called hybrid strategy to resolve this problem. This concept is based on the use of two measures. The first measure aims to analyse the goodness of each partition obtained with different values of input parameters. The use of the second measure permits to select the best partition between those having good but very close values of the first measure. To illustrate this strategy, we propose a new hybrid measure —called "HS-measure"— based on Homogeneity degree and Silhouette coefficient. The performance of our measure is then tested on road traffic data set.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data mining*; I.5.3 [**PATTERN RECOGNITION**]: Clustering

General Terms

Performance

Keywords

clustering, validity measures, hybrid measure

1. INTRODUCTION

Road traffic monitoring systems, generally used by engineers of traffic, permit to supervise traffic state in real time and to propose actions in order to facilitate mobility. Some of them are connected to information systems in order to provide the travellers with information about fluidity of traffic. To offer the functionalities underlined above, these systems use various types of road traffic data: road traffic measures recorded by sensors, images or videos, road users investigations, etc. The collected data has to be analysed and interpreted in order to guarantee suitable solutions to congestions or atypical situations such as accidents, breakdown of traffic lights synchronization,...

The use of knowledge discovery techniques and precisely the clustering process represents an effective tool kit to make such analyses. Indeed, clustering permits to group data having similar characteristics in a same set, called cluster and then to use measures and techniques to validate and interpret the obtained results. In [11], we proposed a new supervised validity measure called *homogeneity degree* that is able to validate only the interpretable results. Applied to road traffic measures, we showed how the homogeneity degree permits to detect typical traffic situations and how these situations can be used to forecast the traffic situation of a day for example or to detect atypical situations (ex. identify atypical fluidity or atypical congestion).

The homogeneity degree can be used as relative measure to find the best clustering results. The use of a relative measure consists on varying an input parameter of the clustering method, that is generally the clusters number, and computing the value of the relative measure for each partition obtained with each value of the input parameter. Then, with most measures we have to select the partition that generates the maximal/minimal value of the relative measure. However, sometimes, the optimal value can be too close to the other generated values so that the choice of this value as the best one becomes meaningless for the user. So, the choice of an optimal value becomes a challenge for the experts and this situation raises doubts about the performance of the evaluation process. A solution to this problem can be the use of a second measure that decides between these close values. To our knowledge, no work has been proposed in this context. In this paper, we propose a new concept called hybrid strategy that illustrates this idea and we present a new measure that applies this strategy that we call —"HS-measure"—. *HS-measure* is based on two measures : the **H**omogeneity degree (a supervised measure) and the **S**ilhouette coefficient (an unsupervised measure). The role of the homogeneity degree is to evaluate the goodness of each partition obtained with different values of input parameters. The use of silhouette coefficient permits to select the best partition between the ones that have close values to the optimal value of homogeneity degree.

VALUETOOLS 2011, May 16-20, Paris, France
Copyright © 2011 ICST 978-1-936968-09-1
DOI 10.4108/icst.valuetools.2011.245713

This paper is organized as follows. Section 2 discusses some related works. Section 3 presents the concept of hybrid strategy and a description of HS-measure. Section 4 presents the comparison between homogeneity degree and HS-measure using road traffic data set. Section 5 concludes the paper.

2. RELATED WORKS

In the literature, there exists several methods that aim to discover the best clustering partition. Some of them are based on combining multiple partitions generated by different clustering algorithms into a single clustering result that represents a consensus between all the generated partitions. This technique is called *cluster ensemble* [5, 6]. Others solutions consist on using *validity measures* [1, 2, 3, 4, 7, 8, 10, 11, 16, 13, 17, 14, 15] and especially the *relative* ones to find the best result. A relative measure is generally an unsupervised or supervised measure that permits to compare different results of clustering. The result that optimizes the measure or that produces a significant change is selected.

Generally, the cluster ensemble is used when the user hasn't any guidelines to select the best result among the generated ones. That is not the case of relative measures since the use of such method is guided by a particular need of the user (e.g. the compactness of the clusters, the association of a class label to each cluster ...). In this paper, we are interesting in this latter method because the need is quite clear : select the best clustering result that assures an interpretability of road traffic activity.

Several relative validity measures were proposed such that silhouette coefficient [8], Dunn [4], Davies Bouldin [3] that are unsupervised measures, and entropy [15], F-measure [2, 17], Rand statistic [13] that are supervised ones. Recently, we proposed a new supervised measure called homogeneity degree [11] and we showed that it has the best performances in comparison to the other measures of the same category, and especially in the interpretation task.

According to [7], the use of relative measures is possible only when the clusters number is one of the input parameters. Based on this, the authors of [7] propose to follow three steps : (1) vary the clusters number k between two predefined numbers *kmin* and $kmax$, (2) for each value of k, apply the clustering algorithm r times by varying the other input parameters, and (3) select the best value of the measure obtained for each k. To select the best partition, there exists two approaches that depend on the choice of the measure. If the measure doesn't exhibit an increasing or a decreasing when the number of clusters increases, then the maximal (or the minimal) value of the measure is chosen. Otherwise, we have to plot the values of the measure as the function of k and select the value of k that generates a significant local change that has the shape of ”knee”.

However, in some cases, the best values obtained with a measure (the highest or the lowest values in the first approach, or those that generate the shape of knee in the second approach) may be very close from each other. For instance, let M be a measure that doesn't exhibit any increasing or decreasing when the number of clusters increases (first approach) and suppose that values domain of this measure is the range [0,1] and that a high value of this measure indicates a good partition. Suppose that we obtain the values of 0.91 and 0.92 as the highest values of M. The difference between these values is too small that the choice of the optimal value becomes meaningless for the user. A solution to this problem can be the use of a second measure that decides between these close values. To our knowledge, no work has been proposed in this context. In this paper, we propose a new concept —called *hybrid strategy*— that is based on this idea. The following section presents this concept.

3. A HYBRID EVALUATION STRATEGY CONCEPT

A hybrid strategy is based on the use of two measures M_1 and M_2. The first measure M_1 is applied on all the partitions generated by varying the number of clusters k. Basing on a *closeness* concept, we select the set of partitions SP that are associated to the best values of M_1 and such that these values are *close* to each other. Then, the second measure M_2 is applied on each partition of SP. The partition that optimizes M_2 and *improves* the results is selected.

The concepts of *closeness* and *improvement* are very related to the choice of the measures M_1 and M_2. To illustrate these concepts, we present in this paper a new measure based on the hybrid strategy that we call *HS-measure.*

3.1 HS-measure : a hybrid clustering validity measure

HS-measure (**H**omogeneity degree and **S**ilhouette coefficient based **measure**) is a relative measure that represents a combination of a supervised measure, that is homogeneity degree, and an unsupervised measure, that is silhouette coefficient. In fact, we think that it is interesting to apply an unsupervised measure on the partitions associated to the best values generated by the supervised measure (and that are very close to each other) in order to choose, so far as we can, the one composed of the more compact and/or the well separated clusters.

The following subsections recall the basic idea of *homogeneity degree* and *silhouette coefficient* and then present the HS-measure-based algorithm that illustrates the *HS-measure.*

3.1.1 Homogeneity degree

In [11], we proposed a new supervised validity measure that we called *homogeneity degree.* The basic idea of this measure is to validate only interpretable clusters according to class labels. We proposed two degrees : the *overall homogeneity degree* that evaluates the validity of each cluster, and the *partition homogeneity degree* that is associated to the partition.

The value of *overall homogeneity degree* of a cluster depends on the extent to which the cluster satisfies the following rules: the *domain majority rule* and the *cluster majority rule.* The satisfaction of the first rule requires the existence of at least one label such as the number of objects described by this label in the studied cluster represents a proportion strictly greater than 50% of all objects associated to the same label. The degree that illustrates this idea is called *homogeneity degree in respect to the domain.* The second rule is very related to the first one since it requires that the number of all objects associated to the labels identified at the first rule represents a proportion strictly greater than the half of cluster cardinality. This rule is illustrated by *partial homogeneity degree*. A cluster verifying these two rules is called *valid-interpretable* cluster. The higher *the homogeneity degree in respect to the domain* and *the partial homogeneity degree* are, the better the *overall homogeneity of a cluster* is.

The definition of *overall homogeneity degree* requires the use of two thresholds $\alpha \in]0.5, 1]$ and $\beta \in]0.5, 1]$ assuring the coherence of results.

We present hereafter, some useful notations and then we describe the intermediate degrees that are mandatory to the computation of *overall homogeneity degree* and the *partition homogeneity degree*

- $X = \{O_1, O_2, \ldots, O_n\}$: a set of n objects, also called *the domain*.
- $L = \{l_1, l_2, \ldots, l_m\}$: a set of m labels.
- $P_k(C_1, C_2, \ldots, C_k)$: a partition of k clusters.
- $SC_{ij} = \{O_r \in C_i : Label(O_r) = l_j\}$: the set of objects of cluster C_i having the label l_j.
- $S_j = \{O_r \in X : Label(O_r) = l_j\}$: the set of objects of X having the label l_j.
- $\| \cdot \|$: set cardinality symbol.

Membership degree of a label to the domain: the membership degree $MD_\alpha(l_j, C_i)$ of a label l_j to the domain X in respect to a cluster C_i expresses the proportion of objects of C_i described by label l_j in respect to the total number of objects of X having the label l_j. $MD_\alpha(l_j, C_i)$ is defined as follows:

$$MD_\alpha(l_j, C_i) = \begin{cases} \frac{\|SC_{ij}\|}{\|S_j\|}, & \text{if } \frac{\|SC_{ij}\|}{\|S_j\|} \geq \alpha \text{ and } \| S_j \| \neq 0; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

As it is shown in Eq. 1, $MD_\alpha(l_j, C_i)$ will be equal to 0 any time the proportion of objects described by label l_j is strictly less than a given threshold α. This threshold represents a sensibility indicator assuring that C_i contains at least $(100 * \alpha)\%$ of objects of X described by the label l_j.

Homogeneity degree of a cluster in respect to the domain: the homogeneity degree of C_i in respect to the domain, denoted $HD_\alpha(C_i)$, represents the average of the non-null membership degree of the different labels to the domain. It is computed as follows:

$$HD_\alpha(C_i) = \begin{cases} \frac{1}{B} \cdot \sum_{j=1}^{m} MD_\alpha(l_j, C_i), & \text{if } B \neq 0; \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where $B = \| \{l_r : l_r \in L \wedge MD_\alpha(l_r, C_i) > 0\} \|$, that is, the number of labels l_r such that $MD_\alpha(l_r, C_i)$ is strictly positive.

A value of $HD_\alpha(C_i) \neq 0$ means that the cluster C_i verifies the *domain majority rule*.

Membership degree of a label to a cluster: the membership degree $MC_\alpha(l_j, C_i)$ of a label l_j to a cluster C_i reflects the proportion of objects of C_i described by l_j. In other words, $MC_\alpha(l_j, C_i)$ permits to measure the importance of a label l_j to a cluster C_i. Formally, $MC_\alpha(l_j, C_i)$ is computed as follows:

$$MC_\alpha(l_j, C_i) = \begin{cases} \frac{\|SC_{ij}\|}{\|C_i\|}, & \text{if } MD_\alpha(l_j, C_i) \geq \alpha; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

As it is shown in Eq. 3, only labels for which $MD_\alpha \geq \alpha$ are included in the definition of $MC_\alpha(l_j, C_i)$. This ensures that only the labels identified by the *domain majority rule* are considered in the computing of $MC_\alpha(l_j, C_i)$.

Partial homogeneity degree of a cluster: the partial homogeneity degree of a cluster C_i, denoted $HC_{\alpha,\beta}(C_i)$, computes the proportion of objects of C_i described by labels that have a certain importance in respect to the domain ($MD_\alpha(l_j, C_i) \geq \alpha$). Like for MD_α, we use a threshold β to force the cluster to contain at least a proportion greater than β. Formally, the partial homogeneity degree $HC_{\alpha,\beta}(C_i)$ is given by Eq. 4:

$$HC_{\alpha,\beta}(C_i) = \begin{cases} \sum_{j=1}^{m} MC_\alpha(l_j, C_i), \\ \text{if } \sum_{j=1}^{m} MC_\alpha(l_j, C_i) \geq \beta; \\ 0, \\ \text{otherwise.} \end{cases} \quad (4)$$

If $HC_{\alpha,\beta}$ of a cluster C_i is $\neq 0$, this implies that the cluster C_i verifies the *cluster majority rule.*

Overall homogeneity degree of a cluster: the overall homogeneity degree of a cluster C_i takes into account the homogeneity degree of C_i in respect to the domain and the partial homogeneity degree of C_i. It is denoted by $D_{\alpha,\beta}(C_i)$ and is computed through Eq. 5 hereafter:

$$D_{\alpha,\beta}(C_i) = HD_\alpha(C_i) \cdot HC_{\alpha,\beta}(C_i). \quad (5)$$

The values domain of overall homogeneity degree of a cluster is $\{0\} \bigcup]0.25, 1]$, where 1 indicates that the cluster is a *fully valid-interpretable* cluster and 0 indicates that the cluster is not a *valid-interpretable* one. A value between 0.25 and 1 indicates the degree of *validity-interpretability* of the cluster.

Partition homogeneity degree : let $P_k(C_1, C_2, \ldots, C_k)$ be a partition of k clusters. The homogeneity degree of P_k is defined by the function $DP_{\alpha,\beta}(P_k)$ and is computed as follows:

$$DP_{\alpha,\beta}(P_k) = \begin{cases} \frac{\sum_{i=1}^{i=k} D_{\alpha,\beta}(C_i)}{k}, & \text{if } \forall i, D_{\alpha,\beta}(C_i) \neq 0; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The domain values of $DP_{\alpha,\beta}(P_k)$ is $\{0\} \bigcup]0.25, 1]$. When $DP_{\alpha,\beta}(P_k)$ is equal to 0, the partition is not *valid-interpretable.* When it is equal to 1, the partition is said *fully valid-interpretable.* For the other values, we propose a decomposition of the range that is based on the semantic meaning of partition homogeneity degree. The range]0.25, 1[is divided in 3 sub-ranges :

- $]0.25, 0.56[$: a partition having a $DP_{\alpha,\beta}$ value belonging to this sub-range is said *weakly valid-interpretable.*
- $[0.56, 0.81[$: a partition having a $DP_{\alpha,\beta}$ value belonging to this sub-range is said *highly valid-interpretable.*
- $[0.81, 1[$: a partition having a value of $DP_{\alpha,\beta}$ belonging to this sub-range is said *strongly valid-interpretable.*

3.1.2 Silhouette coefficient

The silhouette coefficient [8] is an unsupervised clustering validity measure that is based on the compactness and separation concepts. It is associated to every object, every cluster and every partition. To compute the silhouette coefficient of an object O_i, three steps are necessary:

- Compute the average distance a_i between the object O_i and the other objects belonging to the same cluster as O_i. Let C_i be this cluster.
- For every cluster C_j different from C_i, compute the average distance between the objects of C_j and O_i. Let b_i be the lowest average distance.
- The silhouette coefficient of the object O_i is $Sil(O_i) = \frac{b_i - a_i}{\max(a_i, b_i)}$

The silhouette coefficient $Sil(C)$ of a cluster C is equal to the average value of the silhouette coefficients of the objects belonging to this cluster. The silhouette coefficient $Sil(P)$ of a partition P is equal to the average value of the silhouette coefficients of the clusters of P. The value of $Sil(P)$ varies in the range $[-1, 1]$. According to experiences realized by Kaufman and al in [8], when a partition has a value in the range [-1,0.25], it is considered as a non valid partition. When this value belongs to the range]0.25,5], the clusters contain a considerable noise. When it belongs to the range]0.5,0.7], it means that the objects are clearly assigned to each cluster. If it belongs to]0.7,1], it indicates that the clusters are compact and well-separated. We call these 4 ranges in the rest of this paper the *definition ranges* of silhouette coefficient and we note them *range I*, *range II*, *range III* and *range IV*, respectively.

3.2 HS-measure-based algorithm

The principal role of HS-measure is to identify the best partition P^* that maximizes silhouette coefficient among those having the *best values* of homogeneity degree. The silhouette coefficient is applied only if the *best values* of homogeneity degree are very *close.* Moreover, the result obtained by silhouette coefficient is selected if and only if it *improves* the results of homogeneity degree. So, we have to define the concepts of *closeness* between values and *result improvement.* This section presents these two concepts and the HS-measure-based algorithm.

3.2.1 Closeness concept

We propose two techniques to identify close values of partition homogeneity degree:

Range-based technique: in order to define closeness between values, the range-based technique uses the sub-ranges described on the subsection 3.1.1 (see the *Partition homogeneity degree* definition). Therefore, two values V_1 and V_2 are said *close* if they belong to the same sub-range. The exceptional value 1 that can take the values V_1 and V_2 is included in the category *strongly valid-interpretable.*

Threshold-based technique: two values of partition homogeneity degree V_1 and V_2 are said *close* if the difference between these values is less or equal than a threshold Δ ($|V_1 - V_2| \leq \Delta$). We fixed the value of Δ to 0.1 because we consider that this value is pretty low in comparison to the values domain of the partition homogeneity degree (the values domain is $\{0\} \bigcup]0.25, 1]$).

The choice of one of these two closeness techniques depends on the tolerance degree of the user. If the most important for him is to remain in the same category of validity, he must use the range-based technique. However, if he is less tolerant, he can use the threshold-based technique.

3.2.2 Results improvement by silhouette coefficient

Let P_i and P_j be the two partitions selected according to closeness concept and suppose that P_i is the best partition obtained with $DP_{\alpha,\beta}$ ($DP_{\alpha,\beta}(P_i) > DP_{\alpha,\beta}(P_j)$). Let $Sil(P_i)$ and $Sil(P_j)$ be the silhouette coefficients for the partitions P_i and P_j, respectively. The partition P_j is selected as the best partition P^* if:

- $Sil(P_j)$ is greater than $Sil(P_i)$ and
- $Sil(P_j)$ and $Sil(P_i)$ don't belong to the same *definition range* of silhouette coefficient (range I, II, III and IV).

If these two conditions are satisfied, we can consider that the use of silhouette coefficient **improves** the result.

For instance, suppose that $DP_{\alpha,\beta}(P_i) = 0.9$, $DP_{\alpha,\beta}(P_j) = 0.84$, $Sil(P_i) = 0.6$ and $Sil(P_j) = 0.85$. Since $Sil(P_j) > Sil(P_i)$, and since $Sil(P_j)$ belongs to the range IV and $Sil(P_i)$ belongs to the range III, we can consider that the silhouette coefficient improves the result and the partition P_j is selected as the best partition P^* instead of P_i.

```
Function HS_measure( X : Set of objects, L : Set of labels) : Partition
    $r \leftarrow \| L \|$
    [A table for storing the $DP_{\alpha,\beta}(P_k)$ values]
    $T\_DP : array\ [1..r-1]$
    [A table for storing all the generated partitions ($P_k$)]
    $T\_P_k : array\ [1..r-1]$
    [A table for storing the silhouette coefficient values]
    $T\_Silhouette : array\ [1..r-1]$
    $nbelem \leftarrow 0$

    [Compute the Homogeneity degree for each partition and search the closest values]
    For $k$ from 2 to $r$ do
        $P_k \leftarrow$ CLUSTERING(X,k)
        $nbelem \leftarrow nbelem + 1$
        [Compute the Homogeneity degree for each partition $P_k$ and store it in $T\_DP$]
        $T\_DP[nbelem] = DP_{\alpha,\beta}(P_k)$
        [Store the partition $P_k$ in $T\_P_k$]
        $T\_P_k[nbelem] = P_k$
    end For
    F $\leftarrow$ SEARCH_CLOSE_VALUES($T\_DP$,$T\_P_k$, nbelem)

    [Compute the Silhouette coefficient values for the $F$ partitions]
    For $i$ from 1 to $F$ do
        $P_k \leftarrow T\_P_k[i]$
        $T\_Silhouette[i] \leftarrow Sil(P_k)$
    end For

    [Verify the improvement condition]
    indice_max $\leftarrow$ SEARCH_MAX_SILHOUETTE($T\_Silhouette$,$F$)
    check $\leftarrow$ VERIFY_RESULT_IMPROVEMENT(
    $T\_Silhouette[indice\_max]$, $T\_Silhouette[1]$)
    If (check=true) then
        return $T\_P_k[indice\_max]$
    else
        return $T\_P_k[1]$
    end If
End
```

Algorithm 1: HS-measure-based algorithm

3.2.3 Algorithm

The computation of HS-measure requires two steps. The first step consists of computing the homogeneity degree $DP_{\alpha,\beta}(P_k)$ for different values of the clusters number k, and then selecting the highest value of homogeneity degree and the values close to this value according to one of the proposed closeness techniques. Let F be the number of partitions selected according to this technique. The second step proceeds by computing the value of silhouette coefficient for every partition among the F ones and then choosing the one that maximizes silhouette coefficient and improves the results. The algorithm 1 illustrates these two steps[1]. The function CLUSTERING is one of clustering methods having the number of clusters k as input parameters [9, 8, 12]. The function CLOSENESS-TECHNIQUE can be one of the two closeness techniques presented above.

4. EXPERIMENT AND DISCUSSION

The objective of this section is to show how our hybrid measure permits to improve the evaluation of clustering results. The used data set represents road traffic data recorded from sensors placed on roads.

4.1 Road Traffic data set

Road traffic measures permit to describe the state of the traffic in space and time. In this paper, we are interested in flow measure which corresponds to the number of vehicles which pass in a point x in road network during a time interval I. It is expressed in vehicles per unit of time (generally hours or minutes).

More than 300 sensors are placed on a road of a french city. Every day, each sensor records 480 values of flow measure (one value every 3 minutes). An object of 480 attributes $\{A_1, A_2, \ldots, A_{480}\}$ is associated to each sensor where A_1 is the number of vehicles on 00h00, A_2 is the number of vehicles on 00h03, and so on. Along the year 2003, about 365 objects were collected for every sensor. We focus on the data sets given by three of these sensors to experiment our approach : the sensor A, the sensor B and the sensor C. Each of the three data sets contains 365 objects but some of these objects have missing values. Since the proposed algorithm do not deal with missing values, the data sets of sensors A, B and C will contain 228, 229 and 235 values, respectively. The set of labels L associated to these data sets is built from 4 attributes :

- Attribute A_{481} which is associated to the day of a week. Its values domain is $\{1, 2, \ldots, 7\}$.
- Attribute A_{482} indicating if a day is a holiday or not. The values domain of A_{482} is $\{Y, N\}$.
- Attribute A_{483} which is associated to the month. Its values domain is $\{1, 2, \ldots, 12\}$.
- Attribute A_{484} indicating if the day belongs to school holidays or not. The values domain of A_{484} is $\{Y, N\}$.

Instead of creating labels using cartesian product of attributes, we will look through objects and extract only the values of the attributes that appear together. For instance, some values such that $3Y12N$ (a wednesday holiday of December and not belonging to school holidays) or $3Y2Y$ (a wednesday holiday of February and belonging to school holidays) are not associated to any object in data set for the year 2003. So, they are not selected as labels.

4.2 Application of HS-measure-based algorithm

Since our objective is not the comparison of clustering algorithms, we choose the well-known K-means [9] to test our algorithm. The tables 1, 2 and 3 show the results of the algorithm application on data sets of sensors A, B and C, respectively. The following subsections describe these results.

4.2.1 Data set of sensor A

The application of partition homogeneity degree on data set of sensor A generates four valid-interpretable partitions. Among these partitions, P_5 is the one that maximizes the

[1] The complexity of the algorithm depends on the complexity of CLUSTERING method.

k	$DP_{\alpha,\beta}$	Sil
2	0.66	**0.89**
3	0.73	0.42
4	0.74	0.27
5	**0.76**	0.23
6	0	-
7	0	-
8	0	-
9	0	-
10	0	-
...	...	-

Table 1: Application of HS-measure on road traffic data for the sensor A

partition homogeneity degree. According to the *range-based technique* and *threshold-based technique*, P_2, P_3 and P_4 are all selected as the partitions close to P_5 (all these partitions are *highly valid-interpretable* and the difference between the homogeneity degree of these partitions and the homogeneity degree of P_5 is less or equal than 0.1). For the partition P_5, the value of silhouette coefficient is equal to 0.23. For the partition P_2, the value of this coefficient is equal to 0.89. These values satisfy the condition of *results improvement* since 0.89 belongs to the range IV and 0.23 belongs to the range *I*. So the partition P_2 is selected as the best one instead of P_5. This result is completely different from the one generated by the application of homogeneity degree separately. Moreover, the silhouette coefficient value of P_2 is close to 1, which implies that the clusters belonging to P_2 are compact and well separated, comparing to those of P_5 which are not valid. This example shows how our hybrid measure improves the evaluation results.

4.2.2 Data set of sensor B

Table 2 shows the results of HS-measure application on objects associated to the sensor B. In this table, if we focus on the results obtained by the homogeneity degree, we can see that the partition P_4 is the best one since it is associated to the highest value of homogeneity degree. However, according to the *range-based technique*, the value of $DP_{\alpha,\beta}(P_4)$ is close to those generated by P_2, P_3, P_5 and P_6 (all these partitions are *strongly valid-interpretable*). When we apply the silhouette coefficient on the five partitions, the partition P_2 is picked out because its silhouette coefficient improves the results. Therefore, if we didn't use the hybrid measure HS-measure and if we applied only the homogeneity degree, the partition P_4 would be selected as the best one while the difference between $DP_{\alpha,\beta}(P_2)$ and $DP_{\alpha,\beta}(P_4)$ is as low as 0.019. This case also shows the performance of our hybrid strategy.

We can note that even if we used *threshold-based technique*, the partition P_2 would be selected as the best one by HS-measure. The only difference between the application of *threshold-based technique* and *range-based technique* in this example is that the first technique would choose only the values associated to partitions P_2, P_3 and P_5 as close values to $DP_{\alpha,\beta}(P_4)$ because the difference between $DP_{\alpha,\beta}(P_6)$ and the $DP_{\alpha,\beta}(P_4)$ is greater that 0.1.

4.2.3 Data set of sensor C: an example of no improvement

k	$DP_{\alpha,\beta}$	Sil
2	0.93	**0.30**
3	0.942	0.13
4	**0.949**	0.19
5	0.88	0.12
6	0.83	0.16
7	0	-
8	0	-
9	0	-
10	0	-
...	...	-

Table 2: Application of HS-measure on road traffic data for the sensor B

k	$DP_{\alpha,\beta}$	Sil
2	0.92	**0.30**
3	**0.93**	0.29
4	0.82	0.19
5	0	-
6	0	-
7	0	-
8	0	-
9	0	-
10	0	-
...	...	-

Table 3: Application of HS-measure on road traffic data for the sensor C

Table 3 presents an example where the condition of results improvement is not satisfied. Indeed, according to the homogeneity degree, the partition P_3 is considered as the best one because it has the greatest value of $DP_{\alpha,\beta}$. If we apply the *range-based technique*, the partitions P_i $(i = 2, \ldots, 4)$ are selected for the second step that is the computation of silhouette coefficient. When we measure this coefficient for these partitions, the partition P_2 generates the highest value $(Sil(P_2) = 0.30)$. However, the values $Sil(P_2)$ and $Sil(P_3)$ don't satisfy the conditions of improvement since they belong to the same range of definition. So, the partition P_3 remains the best. This example shows the importance of the results improvement condition use. Indeed, this condition represents the basis of the hybrid strategy since it assures a better quality of clustering results. Moreover, recall that the hybrid strategy is only applied if the first measure M_1 generates close values. So, this strategy must assure a real improvement to explain the choice of another best partition different from the one chosen by the first measure, hence the use of the improvement condition.

5. CONCLUSION

The paper presents a new concept called hybrid strategy that allows to choose between close results of a validity measure. We also propose a new measure based on this strategy, that we call HS-measure. This measure is a combination of the supervised measure : the homogeneity degree, and the unsupervised measure : the silhouette coefficient. The first measure is applied on all partitions generated by varying the clusters number. The second measure permits to select the best partition among those having the highest values of homogeneity degree and such as these values are close to each other. As future directions, we intend to generalize the concept of hybrid strategy in order to use it with all types of measures.

6. REFERENCES

[1] E. Amigo, J. Gonzalo, Artiles, and F. J. Verdejo. A comparison of extrinsic clustreing evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[2] N. Chinchor. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding (MUC4 '92)*, pages 22–29, 1992.

[3] D.-L. Davies and D.-W. Bouldin. Cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(2):224–227, 1979.

[4] J.-C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetica*, 4:95–104, 1974.

[5] X. Z. Fern and W. Lin. Cluster Ensemble Selection. In *Proceedings of the SIAM International Conference on Data Mining*, pages 787–797. SIAM, 2008.

[6] R. Ghaemi, M.-N. Sulaiman, H. Ibrahim, and N. Mustapha. A survey: Clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, (50):636–645, 2009.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

[8] L. Kaufman and P. Rousseeuw. *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons, 1990.

[9] J. Mcqueen. some methods for classification and analysis of multivariate observations. In *5th Berkeley Symp. on Math. Statistics and Probability*, pages 281–298, Berkley, USA, 1967.

[10] M. Meilă. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning (ICML Š05)*, pages 577–584, Bonn, Germany, June 2005.

[11] Y. Naija and K. Sinaoui Blibech. A novel measure for validating clustering results applied to road traffic. In *3rd International Workshop on Knowledge Discovery from Sensor Data (SensorKDD-2009)*, pages 105–113, Paris, France, June 2009.

[12] R.-T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *20th Int. Conf. on Very Large DataBases (VLDB)*, pages 144–155, Santiago, Chile, September 1994.

[13] M. W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[14] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of of the 2007 Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 410–420, Prague, June 2007.

[15] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[16] P.-N. Tan, M. Steinbach, and K. Kumar. *Introduction to Data Mining.* Pearson Addison Wesley, 2005.

[17] C. J. van Rijsbergen. *Information Retrieval (2nd ed.).* Butterworth, 1979.