An Approach to Clustering Models Estimation

Ildar R. Baimuratov, Nataly A. Zhukova
ITMO University
Saint Petersburg, Russia
baimuratov.i@gmail.com, nazhukova@mail.ru

***Abstract*—There are numerous clustering algorithms and clustering quality criteria present at the moment. According to clustering tasks, different criteria can be used as a ground for choosing one or another clustering model. But choosing of clustering quality criteria is in turn not based on any formal approach and depends only on an analyst's intuition. We propose a systematic approach to clustering quality estimation by analyzing the structure of clustering models and extracting the factors of clustering quality. Further, we believe that clustering quality estimation requires estimation of knowledge or information produced in result of clustering. However, we show that existing information criteria can not be used as criteria for choosing clustering models and propose a new information criterion for clustering and examine it experimentally.**

I. Introduction

There are numerous clustering algorithms and clustering quality criteria present at the moment. According to clustering tasks, different criteria can be used as a ground for choosing one or another clustering model. But choosing of clustering quality criteria is in turn not based on any formal approach and depends only on an analyst's intuition. Thus, our first goal is to develop a systematic approach to clustering quality estimation by analyzing the structure of clustering models and finding the factors of clustering quality.

Further, we believe that the systematic approach to clustering quality estimation requires estimation of knowledge or information produced in result of clustering. Though there is a group of information criteria present at the moment, we will show that existing information criteria can not be used as criteria for choosing clustering models, because 1) they need some ideal partition to compare its informativeness with the informativeness of resulting partitions or 2) optimal, according to these criteria, partitions are trivial and, consequently, not intuitively informative nor practically applicable. Therefore, we will propose a new information criterion for clustering which, from the one hand, need not any ideal partitions and, from the other hand, estimate clustering models in such a way that the most preferable models are not trivial.

According to this program, in the first section we will describe clustering models in general in order to extract the clustering model structure. We suppose that clustering models are defined by 1) different measures of distance between objects; 2) different measures of distance between clusters; 3) a predefined number of clusters and 4) different clustering algorithms. In the second section we will extract, according to the general structure of clustering models, a number of clustering quality factors, consider different existing clustering criteria and analyze them with respect to the factors. And finally, in the third section we will consider a group of information criteria and show that existing information measures can not be used for choosing clustering models, then we will define a new information criterion and examine it experimentally.

II. Clustering model structure

Under clustering we consider a process of partitioning of a set of objects X, having m-component feature set, to a set of subsets $C = \{C_1, ..., C_k\}$, where each C_i is called "cluster". The reason of this partitioning is that different objects contained in the same cluster are similar according to their features.

Some particular method of partitioning the set X to the set of clusters C is defined by a clustering model M:

$$C = M(X).$$

We suppose, that some particular method of partitioning the set of objects to the set of clusters is defined by the four factors: 1) a measure of distance between objects $\delta(x_i, x_j)$; 2) a measure of distance between clusters $\Delta(C_i, C_j)$; 3) a number of clusters K and 4) by an algorithm of clustering A. Therefore, we consider the clustering model structure $< \delta, \Delta, A, K >$:

$$M =< \delta, \Delta, A, K >.$$

Let us consider some options for elements of the clustering model structure, except of the number of clusters K as it does not require any specific consideration.

A. Measures of distance between objects

Representation of objects of the original set X, having some m-featured description, as points of the m-dimensional space R^m allows to define a measure of distance $\delta(x_i, x_j)$ between objects x_i and x_j[3], that in turn allows to define cluster as a set of pairs, such that the distance between points of these pairs is less then some value σ:

$$C_i = \{(x_i, x_j) : x_i \in X, x_j \in X, \delta(x_i, x_j) < \sigma\}.$$

Given objects $x, y \in X$ and values x_i, y_i of an i-th feature of the objects x and y, there are following options for calculating the distance $\delta(x, y)$:

- *Euclidean distance* is a geometrical distance in euclidean space:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}.$$

- *Squared euclidean distance* assigns bigger values to objects that are more distant from each other:
$$d_2(x,y) = \sum_i (x_i - y_i)^2.$$
- *Hamming distance* is the sum of differences between coordinates:
$$d_H(x,y) = \sum_i |x_i - y_i|.$$
For the Hamming distance an impact of some single bigger distances is less comparing to the euclidean distance as they are not squared.
- *Chebyshev distance* is useful when you need to differentiate one object from another, if they have different coordinates:
$$d_\infty(x,y) = \max |x_i - y_i|.$$

B. Measures of distance between clusters

A resulting partition of the objects also depends on a method of measuring distance $\Delta(C_i, C_j)$ between clusters C_i and C_j. Given measure $\delta(x,y)$ of distance between objects x and y and a number n_k of objects in a cluster C_K, there are following measures of distance between clusters[1]:

- *Nearest neighbor distance* is the distance between two nearest points from different clusters:
$$D_{\min}(C_i, C_j) = \min\{\delta(x,y) | x \in C_i, y \in C_j\}.$$
- *Farthest neighbor distance* is the distance between two the most distant points from different clusters:
$$D_{\max}(C_i, C_j) = \max\{\delta(x,y) | x \in C_i, y \in C_j\}.$$
- *Distance between centroids.* The centroid μ_k of a cluster C_k is defined as following:
$$\mu_k = \frac{1}{n_k} \sum_{x \in C_k} x.$$
Then we can define the distance between two clusters C_i and C_j as a distance between its centroids μ_i and μ_j:
$$D_\mu(C_i, C_j) = \delta(\mu_i, \mu_j).$$
- *Squared distance between centroids:*
$$D^2_\mu(C_i, C_j) = \delta(\mu_i, \mu_j)^2.$$
- *Mean distance*:
$$D_{mean}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} \delta(x,y).$$

C. Clustering algorithms

Finally, the resulting partition of the objects depends on a clustering algorithm being used. In this research we consider only the k-group clustering algorithms. Let us denote by $d_\mu(C_k)$ the mean distance from points of a cluster C_k to its centroid μ_k:

$$d_\mu(C_k) = \frac{1}{n_k} \sum_{x_i \in C_k} ||x_i - \mu_k||.$$

Given cluster partitioning $C = \{C_1, ..., C_k\}$, there is the least squares function estimating partition quality:

$$SSE(C) = \sum_k d_\mu(C_k) = \sum_k \sum_{x_i \in C_k} ||x_i - \mu_k||^2.$$

Clustering algorithms of this group aims to find by numerous iterations a partitioning C with minimal value of $SSE(C)$:

$$C = arg\min_C \{SSE(C)\}.$$

K-means algorithm[2]:

1) Randomly choose k points as starting centroids.
2) Assign each point to a cluster with the nearest centroid.
3) Recalculate the centroids.
4) Repeat until the centroids will stop changing.

The K-means algorithm depends on starting choices of centroids, therefore, some choices could lead to "bad" partitions.

There is a modification of the k-means algorithm called *k-means++*[4]:

1) Randomly choose k starting centroids.
2) Calculate for each point a distance to the nearest centroid.
3) Choose new centroids with respect to the calculated distances.
4) Continue as in the k-means algorithm.

The advantage of this algorithm is choosing more appropriate points as the starting centroids.

There is a variation of k-means named *k-medoids*[5]. Medoid is an object of the original set chosen as a centroid. K-medoids:

1) Randomly choose k objects as starting centroids.
2) Assign each object to a cluster with the nearest centroid.
3) Recalculate the medoids.
4) Repeat until the medoids stop changing.

This algorithm in combination with the Hamming distance is more resistant to noises in datasets than k-means.

Another variation of k-means is *k-medians*[6], where medians are calculated and chosen as centroids instead of mean values. A median of a cluster is an object with the median coordinates with respect to coordinates of other objects of the cluster. K-medians:

Randomly choose k points as starting centroids.
1) Assign each object to a cluster with the nearest centroid.
2) Recalculate the medians and choose them as new centroids.
3) Repeat until the centroids stop changing.
4)

Further we will consider different existing criteria of evaluating clustering quality.

III. Clustering quality criteria

Under a clustering quality criterion we understand a function defined on a set of partitions of a set of objects, according to which it is possible to choose one partition to another. Under optimal clustering we mean partitioning that has maximal or minimal, depends on the criterion, value of the function. At the moment, choosing of clustering quality criteria is not restricted by any formal system and depends on analyst's experience and intuition[1].

Clustering quality criteria can be divided into two groups. If for calculating of a criterion you use only the set of objects and the resulting partition itself, then it is an *internal* criterion[2][7]. Otherwise, if the result of clustering is compared with some ideal partition, then it is an *external* criterion. According to our task, only internal criteria are considered.

As internal criteria of clustering quality are based only on the original data set and the resulting partition, they should

depend on the clustering model structure described in the previous section. We will suggest a list of clustering quality factors, based on the clustering model structure. It appears, that improving one part of the structure leads to decreasing characteristics of others parts[2], therefore, clustering quality criteria must account all parts of the clustering model structure. Thus, we will consider existing clustering quality criteria and analyze it according to this conclusion.

A. Factors of clustering quality

According to the clustering models structure, we consider the following factors of clustering quality:

- *A number of clusters* K and its derivatives (K-factor). The factor is accounted by a criterion only if it is a coefficient of a criterion, but it is useful to list:
  - a number of points n_k in some cluster C_k;
  - a number of pairs of points p_k of a cluster C_k:
    $$p_k = n_k(n_k - 1)/2;$$
  - a number of within-group pairs p_w for all clusters:
    $$p_w = \sum_k p_k = \frac{1}{2}(\sum_k n_k^2 - N);$$
  - a number of between-group pairs p_b of points which do not belong to the same cluster:
    $$p_b = \sum_{k<k'} = n_k n_{k'}.$$
  - a total number of pairs p_t for all points:
    $$p_t = \frac{N(N-1)}{2}.$$
- *Within-group factor* (WG-factor). Options are:
  - within-group scatter matrix WG_k, defined for each cluster C_k:
    $$WG_k = X_k^t X_k,$$
    where X_k is the matrix formed by the centered vectors $v_k = x_i - \mu_k$ for all $x_i \in C_k$. Also it is useful to define the total within group scatter matrix WG as
    $$WG = \sum_k WG_k;$$
  - within-group dispersion $WGSS_k$ of a cluster C_k, which is a sum of squared distances between points and a centroid of C_k:
    $$WGSS_k = \sum_{x \in C_k} ||x_i - \mu_k||^2.$$
    Also it is useful to define the total within-group dispersion $WGSS$ as a sum of within-group dispersions for all clusters:
    $$WGSS = \sum_k WGSS_k;$$
  - the sum S_W of the p_w distances between all the pairs of points inside each cluster:
    $$S_W = \sum_k \sum_{x_i, x_j \in C_k, i<j} \delta(x_i, x_j);$$
  - the maximal within cluster distance $d_{\max}(C_k)$ for C_k:
    $$d_{\max}(C_k) = \max_{x_i, x_j \in C_k, i \neq j} ||x_i - x_j||$$
    and the maximal within cluster distance $d_{\max}$ for all clusters:
    $$d_{\max} = \max(d_{\max}(C_k);$$
  - the sum E_W of distances from the points of each cluster C_k to its centroid μ_k:
    $$E_W = \sum_k d_\mu(C_k).$$
- *Between-group factor* (BG-factor). Options are:
  - between group scatter matrix BG_k, defined for each cluster k:
    $$BG_k = B^t B,$$
    where B is the matrix formed by the vectors $v = \mu_k - \mu$, being reproduced n_k times. Also it is useful to define a total scatter matrix T:
    $$T = X^t X,$$
    where X is the matrix formed by the vectors $v = x_i - \mu$ for all $x_i \in X$;
  - between-group dispersion $BGSS$, which is a weighted sum of squared distances between centroids of all clusters and a centroid of all points μ:
    $$BGSS = \sum_k n_k ||\mu_k - \mu||^2;$$
  - the sum S_B of between cluster distances:
    $$S_B = \sum_{k<k'} \sum_{i \in C_k, j \in C_{k'}, i<j} \delta(x_i, x_j);$$
  - the minimal between cluster distance:
    $$D_{\min} = \min(D_{\min}(C_k, C_{k'});$$
  - the maximal distance between clusters centroids:
    $$D_{\max \mu} = \max_{k<k'}(D_\mu(C_k, C_{k'});$$

We considered a number of existing internal criteria[7]. The results of analysis are represented in the Table I

Summing up, the only appropriate criteria are the Calinski-Harabasz and the PBM indexes. We chose Calinski-Harabasz (CH) index for further usage.

IV. Information criteria

As partitioning of the set of objects X to the set of clusters C can be considered as probability distribution, different information-theoretical characteristics can be applied to this partitioning to estimate an amount of information, produced in result of clustering.

A. Existing information characteristics

There are following characteristics in information theory at the moment[8]:

- *Self-information* for some value x_i of a discrete random variable X with probability distribution $P(x_i)$:
  $$I(x_i) = -\log P(x_i).$$
- *Entropy* of a discrete random variable X:
  $$H(X) = -\sum_i P(x_i) \log P(x_i).$$
- *Conditional entropy* of discrete random variables X and Y with probability distributions $P(x_i)$ and $P(y_j)$ and joint probability distribution $P(x_i|y_j)$:

[a] $S_{\min}$ is the sum of the p_w smallest distances from all the p_t pairs of points, $S_{\max}$ — the sum of the p_w largest distances from all the p_t pairs of points

[b] s^+ is the number of times $\delta(x, x') < \delta(y, y')$, where $x, x' \in C_i, y \in C_k, y' \in C_{k'}$, s^- is the number if times where the opposite situation occurs.

[c] Formulae are not listed due to complexity of its representation.

TABLE I. INTERNAL CRITERIA ANALYSIS

Criterion	Formula	WG-factor	BG-factor	K-factor
Ball-Hall index	$= \frac{1}{K}\sum_k d_\mu(C_k)$	+	-	-
The Banfeld-Raftery index	$= \sum_k n_k \log(\frac{Tr(WG_k)}{n_k})$	+	-	-
The C-index	$= \frac{S_W - S_{\min}}{S_{\max} - S_{\min}}$ a	+	+	-
The Calinski-Harabasz index	$= \frac{N-K}{K-1}\frac{BGSS}{WGSS}$	+	+	+
The Davies-Bouldin index	$= \frac{1}{K}\sum_k \max_{k' \neq k}(\frac{d_\mu(C_k)+d_\mu(C_{k'})}{D_\mu(C_k,C_{k'})})$	+	+	-
The Det-Ratio index	$= \frac{\det(T)}{\det(WG)}$	+	-	-
The Dunn index	$= \frac{D_{\min}}{d_{\max}}$	+	+	-
The Baker-Hubert Gamma index	$= \frac{s^+ - s^-}{s^+ + s^-}$ b	+	+	-
The GDI index	$= \frac{\min_{k \neq k'} \Delta(C_k, C_{k'})}{\max \delta(C_k)}$	+	+	-
The G plus index	$= \frac{2s^-}{p_t(p_t-1)}$	+	+	-
The $K^2\|W\|$-index	$= K^2 \det(WG)$	+	-	+
The Log_Det_Ratio index	$= N \log(\frac{det(T)}{det(WG)})$	+	-	-
The Log_SS Ratio index	$= \log(\frac{BGSS}{WGSS})$	+	+	-
The McClain-Rao index	$= \frac{p_b}{p_w}\frac{S_W}{S_B}$	+	+	-
The PBM index	$= (\frac{1}{K}\frac{\sum_{x_i \in N}\delta(x_i,\mu)}{E_W} D_{\max \mu})^2$	+	+	+
The Point-Biserial index	$= (\frac{S_W}{p_w} - \frac{S_B}{p_b})\frac{\sqrt{p_w p_b}}{p_t}$	+	+	-
The Ray-Turi index	$= \frac{1}{N}\frac{WGSS}{\min_{k<k'}(D_\mu^2(C_k,C'_k))}$	+	+	-
The Scott-Symons index	$= \sum_k n_k \log(\det(\frac{WG_k}{n_k}))$	+	-	-
The SD index	c	+	+	-
The Silhouette index	c	+	+	-
The Tau index	$= \frac{s^+ - s^-}{\sqrt{p_b p_w (\frac{p_t(p_t-1)}{2})}}$	+	+	-
The Tr(W) index	$= WGSS$	+	-	-
The Tr($W^{-1}B$) index	$= Tr(WG^{-1}BG)$	+	+	-
The Wemmert-Gancarski index	$= \frac{1}{N}\sum_k \max(0, n_k - \sum_{x_i \in N}\frac{\|\|x_i-\mu_k\|\|}{\min_{k \neq k'}\|\|x_i-\mu_k\|\|})$	+	+	-
The Xie-Beni index	$= \frac{1}{N}\frac{WGSS}{D_{\min}(C_k,C_{k'})^2}$	+	+	-

$$H(Y|X) = -\sum_i \sum_j P(x_i)P(y_j|x_i)\log P(y_j|x_i).$$

- *Relative entropy* (Kullback–Leibler divergence) of discrete random variables X and Y with n values:

$$D_{KL}(X||Y) = \sum_i P(x_i)\log\frac{P(x_i)}{P(y_i)}.$$

There is a group of information criteria amongst external criteria as well. The reason of external information criteria is to compare an amount of information produced in result of clustering with an amount of information contained in an ideal partitioning. These criteria are *Mutual information* $I(X,Y)$ of two discrete random variables X and Y:

$$I(X,Y) = H(Y) - H(Y|X);$$

and adjusted mutual information[9].

The idea of adjusted mutual information is to adjust mutual information of two random variables with respect to a probability of its joint partitions. Given N points and two partitions U and V, the number of points $a_i = |U_i|$ for $U_i \subseteq U, i = 1...R$ and $b_j = |V_j|$ for $V_j \in V, j = 1...C$, the total number of ways to jointly assign N points into two partitions U and V is Ω:

$$\Omega = \frac{(N!)^2}{\prod_i a_i! \prod_j b_j!}.$$

Each joint partition of U and V can be represented as a contingency table M:

$$M = [n_{ij}]_{j=1...C}^{i=1...R}.$$

Given some contingency table M, there are w different ways of assigning the data points, that will result in this particular M:

$$w = \frac{N!}{\prod_i \prod_j n_{ij}!}.$$

Thus, the probability $P(M|a,b)$ of M with respect to a set $\mathcal{M}$ of all possible contingency tables is specified by

$$P(M|a,b) = \frac{w}{W}.$$

Given mutual information $I(M)$ for the contingency table M,

$$I(M) = \sum_i \sum_j \frac{n_{ij}}{N}\log\frac{N\dot{n}_{ij}}{a_i b_j},$$

the average mutual information of all possible joint partitions of random variables X and Y is *expected mutual information* $E(I(M)|a,b)$:

$$E(I(X,Y)) = E(I(M)|a,b) = \sum_{M \in \mathcal{M}} I(M)P(M|a,b).$$

Then *adjusted mutual information* $AI(X,Y)$ is defined as follows:

$$AI(X,Y) = \frac{I(X,Y) - E(I(X,Y))}{max(H(X),H(Y)) - E(I(X,Y))}.$$

Though external information criteria do not fit to our goals as they need some predefined ideal partitioning, the idea of adjusting an amount of information contained in some partition with respect to its probability will be useful.

B. An analysis of the existing information characteristics

First, we are going to show that conditional entropy $H(Y|X)$ in general case can not be used as information criterion for clustering. We denote a number of elements of cluster C_k under condition x_i as n_{ik}:

$$n_{ik} = |X_i \cap C_k|,.$$

where X_i is a set of objects that satisfy the condition x_i. As for non-fuzzy clustering each $x_i \in X$ is assigned to only one cluster C_k, for each i and k holds that $n_{ik} = 1$, therefore, for each C and X the conditional entropy $H(C|X)$ of non-fuzzy clustering is zero:

$$H(C|X) = -\sum_{x_i \in X} \sum_k \frac{1}{N} \log_N 1 = 0.$$

Second, we are going to use the number of points N as base of logarithmic function of information criteria in order to get normalized values. Besides of other advantages of normalized variants, it will be easier to show with them that existing information criteria can not be used for choosing clustering models. We can do so as base of the logarithmic function in the definitions of the criteria are not restricted.

Summing up, we have following information characteristics, applicable to results of clustering:

- Clustering informativeness C:
$$H(C) = -\sum_k \frac{n_k}{N} \log_N \frac{n_k}{N}.$$
- Clustering informativeness C relative to the original set X:
$$D_{KL}(C||X) = \sum_k \frac{n_k}{N} \log_N \frac{\frac{n_k}{N}}{\frac{1}{N}} = \sum_k \frac{n_k}{N} \log_N n_k.$$

Now we are going to show that these information criteria can not be used for choosing clustering models. For this we will compare values of Calinski-Harabasz index (KH) with values of information characteristics $H(C)$ and $D_{KL}(C||X)$ with respect to different numbers of clusters n from 1 to 10. We will use Fisher's Iris flowers as dataset and k-means as clustering algorithm. The results of comparing are in the Table II:

TABLE II. EXISTING INFORMATION CRITERIA COMPARING

K	CH	$H(C)$	$D_{KL}(C\|\|X)$
1	n/a	n/a	1
2	513	0,13	0,87
3	560	0,22	0,78
4	529	0,27	0,73
5	494	0,3	0,7
6	475	0,35	0,65
7	451	0,38	0,62
8	441	0,4	0,6
9	409	0,41	0,59
10	391	0,44	0,56

If we use information criterion $H(C)$ for choosing number of clusters n, we should choose $n = 10$, or if we use D_{KL} — $n = 1$. We generalize this result as following:

- $H(C) = 1 \Leftrightarrow \forall i |C_i| = 1$,
- $D_{KL}(C||X) = 1 \Leftrightarrow C = X$.

In other words, if we use information characteristic $H(C)$ for choosing the number of clusters, then the optimal number of clusters equals to the number of objects, otherwise, if we use information characteristic $D_{KL}(C||X)$, then the optimal partition is the set X itself. Both these options do not correspond to the objective structure of data nor values of clustering quality criteria.

C. A new information criterion

Therefore, we propose a new information criterion for choosing clustering models. The idea of this criterion is similar to the idea of adjusted mutual information, but it does not need predefined ideal partition. This criterion is a conditional entropy with respect to the set of possible partitions $Part(X)$ of the set X.

We define $\mathcal{X}$ as a set of all possible transformations of the set X. The number of all possible transformations is defined as follows:

$$|\mathcal{X}| = |X|^{|X|}.$$

Then for each partition $Part_i(X)$ we define probability $P(Part_i(X))$ with respect to the set of all possible transformations $\mathcal{X}$. Given $|X| = n, |X_i| = n_i, |Part_i(X)| = k$, and a number k_j of all subsets X_i with the same number of elements n_i, the number of transformations $|Part_i(X)|$ for the partition $Part_i(X)$ is defined by the formula

$$|Part_i(X)| = \frac{n!}{n_1!...n_k!} \frac{n!}{k!(n-k)!} \frac{k!}{k_1!...k_m!},$$

therefore, the probability $P(Part_i(X))$ is defined as following:

$$P(Part_i(X)) = \frac{|Part_i(X)|}{|\mathcal{X}|},$$

Now we can define weighted entropy $P(Part(X))H(X)$ and mean weighted entropy $H(X|Part(X))$ of the set X with respect to the set of all possible partitions $Part(X)$:

$$H(X|Part(X)) = \sum_i P(Part_i(X))H_i(X).$$

Beside of this numerical measure, it is possible to define a binary characteristic H^+:

$$H^+ = 1 \Leftrightarrow P(Part(X))H(X) \geq H(X|Part(X))$$

and a measure of difference $d(H(X), H(X|Part(X))$ of entropy $H(X)$ from the mean value $H(X|Part(X))$:

$$d(H(X), H(X|Part(X)) = |H(X) - H(X|Part(X))|.$$

Let us test the proposed measures as the criteria for choosing clustering models by comparing it with Calinski-Harabasz index like we do with characteristics $H(X)$ and $D_{KL}(C||X|)$. We have to take a random data set with less number of elements to avoid overcomplicated calculations. Let it be a random sample of data from Fisher's flowers data set with

TABLE III. CLUSTERING RESULTS

n	x_1	x_2	x_3	x_4	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	5.0	3.2	1.2	0.2	1	0	0	3	0	1	1	7	7
2	5.8	2.7	4.1	1.0	0	2	2	2	5	5	5	5	2
3	6.9	3.2	5.7	2.3	0	1	3	1	3	2	3	3	0
4	6.1	2.6	5.6	1.4	0	1	1	4	1	3	0	0	3
5	7.1	3.0	5.9	2.1	0	1	3	1	3	2	7	6	6
6	4.4	3.2	1.3	0.2	1	0	0	0	4	4	4	4	4
7	5.2	3.4	1.4	0.2	1	0	0	3	0	6	6	1	1
8	4.4	3.2	1.3	0.2	1	0	0	0	4	4	4	4	4
9	5.7	3.6	3.5	1.0	0	2	2	2	2	0	2	2	5
10	5.2	3.4	1.4	0.2	1	0	0	3	0	6	6	1	1

TABLE IV. THE NEW INFORMATION CRITERIA COMPARING

K	CH	$H(X)$	$P(Part(X))H(X)$	$d(H(X), H(X\|Part(X))$	H^+
2	39	0.552827	0.000001	0.321273	1
3	81	0.527096	0.000096	0.295542	1
4	101	0.492621	0.000938	0.261067	1
5	187	0.286272	0.010907	0.054718	1
6	262	0.263548	0.050209	0.031994	1
7	328	0.143136	0.027269	-0.088417	0
8	1	0.013763	0.120411	-0.111142	0
9	1	0.013763	0.120411	-0.111142	0
10	1	0.120411	0.013763	-0.111142	0

10 elements. The data set and results of the clustering are represented in the Table III. The results of criteria calculation are in the Table IV.

Summing up, the mean weighted entropy of the set with 10 elements is $H(X|Part(X)) = 0.231554$, the partition with $K = 6$ has the maximal weighted entropy $P(Part(X))H(X) = 0.050209$, the partitions with $K = 2...6$ are informative with respect to the mean value. The value of the criterion $P(Part(X))H(X)$ slightly differs from the Calinski-Harabasz index: according to Calinski-Harabasz we should choose $K = 7$, but according to our information criterion we should choose $K = 6$. We suppose, the final decision depends on priorities.

V. CONCLUSION

Summing up, we described the general structure of the clustering models and considered some options for its elements. Then we extracted factors of clustering quality according to the general structure, analyzed different existing internal criteria of clustering quality with respect to these factors and chose a criterion that accounts all elements of the clustering models structure. After that we considered different information characteristics and analyzed it with respect to its applicability as the criterion for choosing clustering models. We proved that none of them is applicable and proposed a new criterion. Finally, we examined it experimentally and got successful result.

At the current point, the practical applicability of the proposed criterion is limited as the number of possible transformations of a given set grows extremely fast, but we believe that this limitation can be overcome in further researches.

REFERENCES

[1] Айвазян С. А., Бухштабер В. М., Енюков Е. С. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989.

[2] Zaki M. J., Meira Jr. W., Meira W. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[3] Барсегян А. Методы и модели анализа данных: OLAP и Data Mining. СПб., БХВ-Петербург, 2004.

[4] Arthur D., Vassilvitskii S. "k-means++: The advantages of careful seeding", *in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. – Society for Industrial and Applied Mathematics*, 2007, pp. 1027-1035.

[5] Kaufman L., Rousseeuw P. *Clustering by means of medoids*. North-Holland, 1987.

[6] Jain A. K., Dubes R. C. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[7] Desgraupes B. "Clustering indices", *University of Paris Ouest-Lab Modal'X*. vol.1, 2013, pp. 1-34.

[8] Габидулин Э. М., Пилипчук Н. И. Лекции по теории информации. М.: МИФИ, 2007.

[9] Vinh N. X., Epps J., Bailey J. "Information theoretic measures for clusterings comparison", *Proceedings of the 26th Annual International Conference on Machine Learning - ICML*, 2009, pp. 1073-1080.