$$U(\pi_C, \Pi) = \sum_{i=1}^{H} U(\pi_C, \pi_i). \quad (13)$$

Therefore, the best median partition should maximize the value of overall utility:

$$\pi_C^{\text{best}} = \arg\max_{\pi_C} \; U(\pi_C, \Pi). \quad (14)$$

Importantly, Mirkin [39] has proven that maximization of partition utility in (13) is equivalent to minimization of the square-error clustering criterion if the number of clusters K in target partition π_C is fixed. This is somewhat surprising in that the partition utility function in (14) uses only the between-attribute similarity measure of (12), while square-error criterion makes use of distances between objects and prototypes. Simple standardization of categorical labels in $\{\pi_1, \ldots, \pi_H\}$ effectively transforms them to quantitative features [39]. This allows us to compute real-valued distances and cluster centers. This transformation replaces the ith partition π_i assuming $K(i)$ values by $K(i)$ binary features, and standardizes each binary feature to a zero mean. In other words, for each object x we can compute the values of the new features $\tilde{y}_{ij}(x)$, as following:

$$\tilde{y}_{ij}(x) = \delta(L_j^i, \pi_i(x)) - p(L_j^i), \text{for } j = 1 \ldots K(i), i = 1 \ldots H. \quad (15)$$

Hence, the solution of median partition problem in (4) can be approached by k-means clustering algorithm operating in the space of features $\tilde{y}_{ij}$ if the number of target clusters is predetermined. We use this heuristic as a part of empirical study of consensus functions.

Let us consider the information-theoretic approach to the median partition problem. In this framework, the quality of the consensus partition π_C is determined by the amount of information $I(\pi_C, \Pi)$ it shares with the given partitions in $\prod$. Strehl and Ghosh [47] suggest an objective function that is based on the classical Shannon definition of mutual information:

$$\pi_C^{\text{best}} = \arg\max_{\pi_C} \; I(\pi_C, \Pi), \text{ where } I(\pi_C, \Pi) = \sum_{i=1}^{H} I(\pi_C, \pi_i), \quad (16)$$

$$I(\pi_C, \pi_i) = \sum_{r=1}^{K} \sum_{j=1}^{K(i)} p(C_r, L_j^i) \log \left(\frac{p(C_r, L_j^i)}{p(C_r) p(L_j^i)} \right). \quad (17)$$

Again, an optimal median partition can be found by solving this optimization problem. However, it is not clear how to directly use these equations in a search for consensus.

We show that another information-theoretic definition of entropy will reduce the mutual information criterion to the category utility function discussed before. We proceed from the generalized entropy of degree s for a discrete probability distribution $P = (p_1, \ldots, p_n)$ [23]:

$$H^s(P) = (2^{1-s} - 1)^{-1} \left(\sum_{i=1}^{n} p_i^s - 1 \right), \quad s > 0, \; s \neq 1. \quad (18)$$

Shannon's entropy is the limit form of (18):

$$\lim_{s \to 1} H^s(P) = -\sum_{i=1}^{n} p_i \log_2 p_i. \quad (19)$$

Generalized mutual information between σ and π can be defined as:

$$I^s(\pi, \pi_C) = H^s(\pi) - H^s(\pi | \pi_C). \quad (20)$$

Quadratic entropy ($s = 2$) is of particular interest since it is known to be closely related to classification error, when used in the probabilistic measure of interclass distance. When $s = 2$, generalized mutual information $I(\pi_C, \pi_i)$ becomes:

$$\begin{aligned} &I^2(\pi_C, \pi_i) = \\ &-2\left(\sum_{j=1}^{K(i)} p(L_j^i)^2 - 1 \right) \\ &+2 \sum_{r=1}^{K} p(C_r) \left(\sum_{j=1}^{K(i)} p(L_j^i | C_r)^2 - 1 \right) = 2 \sum_{r=1}^{K} p(C_r) \sum_{j=1}^{K(i)} p(L_j^i | C_r)^2 \\ &-2 \sum_{j=1}^{K(i)} p(L_j^i)^2 = 2\, U(\pi_C, \pi_i). \end{aligned} \quad (21)$$

Therefore, generalized mutual information gives the same consensus clustering criterion as category utility function in (13). Moreover, traditional Gini-index measure for attribute selection also follows from (12) and (21). In light of Mirkin's result, all these criteria are equivalent to within-cluster variance minimization, after simple label transformation. Quadratic mutual information, mixture model, and other interesting consensus functions have been used in our comparative empirical study.

5 Combination of Weak Clusterings

The previous sections addressed the problem of clusterings combination, namely, how to formulate the consensus function regardless of the nature of individual partitions in the combination. We now turn to the issue of generating different clusterings for the combination. There are several principal questions. Do we use the partitions produced by numerous clustering algorithms available in the literature? Can we relax the requirements for the clustering components? There are several existing methods to provide diverse partitions:

1. Use different clustering algorithms, e.g., k-means, mixture of Gaussians, spectral, single-link, etc., [47].
2. Exploit built-in randomness or different parameters of some algorithms, e.g., initializations and various values of k in k-means algorithm [35], [15], [16].
3. Use many subsamples of the data set, such as bootstrap samples [10], [38].

These methods rely on the clustering algorithms, which are powerful on their own and as such are computationally involved. We argue that it is possible to generate the partitions using weak, but less expensive, clustering algorithms and still achieve comparable or better performance. Certainly, the key motivation is that the synergy of many such components will compensate for their weaknesses. We consider two simple clustering algorithms:

1. Clustering of the data projected to a random subspace. In the simplest case, the data is projected