A Probabilistic Model Using Information Theoretic Measures for Cluster Ensembles

Hanan Ayad, Otman Basir, and Mohamed Kamel

Pattern Analysis and Machine Intelligence Lab
Systems Design Engineering, University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
{hanan,mkamel}@pami.uwaterloo.ca
http://pami.uwaterloo.ca/

Abstract. This paper presents a probabilistic model for combining cluster ensembles utilizing information theoretic measures. Starting from a co-association matrix which summarizes the ensemble, we extract a set of *association distributions*, which are modelled as discrete probability distributions of the object labels, conditional on each data object. The key objectives are, first, to model the associations of neighboring data objects, and second, to allow for the manipulation of the defined probability distributions using statistical and information theoretic means. A Jensen-Shannon Divergence based Clustering Combination (JSDCC) method is proposed. The method selects cluster prototypes from the set of association distributions based on entropy maximization and maximization of the generalized JS divergence among the selected prototypes. The method proceeds by grouping association distributions by minimizing their JS divergences to the selected prototypes. By aggregating the grouped association distributions, we can represent empirical cluster conditional probability distributions of the object labels, for each of the combined clusters. Finally, data objects are assigned to their most likely clusters, and their cluster assignment probabilities are estimated. Experiments are performed to assess the presented method and compare its performance with other alternative co-association based methods.

1 Introduction

Unsupervised classification, or data clustering is an essential tool for exploring and searching for groups in unlabelled data. It is a challenging problem because the clusters inherent in the data can be of arbitrarily different shapes and sizes. A large number of clustering techniques have been developed over the years [1, 2]. However, many are limited to finding clusters of specific shapes and structures and may fail when the data reveal cluster shapes and structures that do not match their assumed model. For instance, the K-means which is one of the simplest and computationally efficient clustering technique, can easily fail if the true clusters inherent in the data are not hyper-spherically shaped.

Recently, there has been an emergent interest in studying cluster ensembles to enhance the quality and robustness of data clustering and to accommodate a

F. Roli, J. Kittler, and T. Windeatt (Eds.): MCS 2004, LNCS 3077, pp. 144–153, 2004.

wider variety of data types and clusters structures [3–10]. Some of the research have relied on using a co-association matrix as a voting medium for finding the combined partitioning. Fred and Jain [5], used single link (SLink) hierarchical clustering to produce a final partitioning of the data. Cluster-based Similarity Partitioning Algorithm (CSPA) is a consensus function introduced by Strehl and Ghosh [3] in which graph partitioning is applied to the co-association matrix resulting in a consensus partitioning. In [8, 9], we used the co-association matrix to construct a Weighted Shared nearest neighbors Graph (WSnnG) and applied a weighted version of the same graph partitioning software tool METIS [11] used in CSPA to find the consensus partitioning.

In this paper we present a probabilistic model derived from the co-association matrix, in which each object's co-associations are modelled as a probability distribution, which we refer to as association distribution. The model is presented in Section 2. A proposed information theoretic clustering combination process, which generates groups of association distributions is presented in Section 3. Groups of association distributions are aggregated through averaging to represent empirical cluster conditional probability distributions for each of the combined clusters. The model allows for cluster assignment probabilities to be estimated. Experimental results and analysis are presented in Section 4. Five datasets with various cluster structures are used to evaluate the performance of the proposed method at different values of the design parameters and in comparison with alternative methods that operate on the co-association matrix.

It is noted that diversity among the data clusterings of the ensemble can be created in a number of different ways, such as random restarts of a single clustering technique, or the use of different clustering techniques, or data resampling as has been reported recently in [12, 13]. In this paper, we use multiple K-means clusterings using random initial restarts.

2 Probabilistic Model for Cluster Ensembles

2.1 Problem Formulation

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a set of n data objects described as d-dimensional vectors of features. Let $\{x_1, \cdots, x_n\}$ denotes a set of n labels corresponding to each of the data objects. Let a cluster ensemble consists of m data clusterings of the n d-dimensional data vectors $\{\mathbf{x_i}\}_{i=1}^n$. While the clusterings are performed on the objects as vectors in their d-dimensional feature space, the modelling and combination methods described in this work deals exclusively with object labels, rather than objects as feature vectors. Throughout the paper, we will use the terms objects and object labels interchangeably to refer to object labels denoted by $\{x_i\}_{i=1}^n$, unless it is otherwise specified.

Let $\{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ be m n-dimensional labelling vectors representing the data clusterings where each clustering $\mathbf{y}_i$ of the n objects consists of a number k_i of clusters. Each entry y_{ij} in the vector $\mathbf{y}_i$ represents the cluster label, i.e. index of the cluster, to which data object x_j is assigned, such that $y_{ij} \in \{1, \cdots, k_i\}$. In this paper, the m clusterings are combined resulting in a clustering $\mathbf{y}$ which consists

of k^* clusters where k^* is prespecified. In addition, an n-dimensional probability vector is generated which gives estimate of the probability of association of each object to its assigned cluster.

2.2 Association Distributions

Let $\mathbf{S}$ be an $n \times n$ co-association matrix which summarizes the generated ensemble. Each entry s_{ij} of $\mathbf{S}$ represents the ratio of the number of times objects x_i and x_j co-occur in the same cluster to the number of clusterings m.

Let X be a discrete random variable which takes as values the object labels $\{x_1, \cdots, x_n\}$. Given an object label x_i, define an association distribution $p(x|x_i)$ as a probability distribution of the random variable X. We have a total of n association distributions, given each object label x_i. The probability values assumed by $p(x|x_i)$ are computed as follows:

$$P(X = x_j|x_i) = \frac{s_{ij}}{\sum_{k=1}^{n} s_{ik}} \qquad \forall j \in \{1, \cdots, n\}$$

That is, each association distribution is simply computed by normalizing each row/column of $\mathbf{S}$. Hence, $p(x|x_i)$ satisfies the two conditions: $P(X = x_j|x_i) \geq 0, \forall j \in \{1, \cdots, n\}$, and $\sum_{j=1}^{n} P(X = x_j|x_i) = 1$. Each data object x_i contributes $\frac{1}{n} p(x|x_i)$ to the estimated probability distribution $p(x)$ of X, i.e.,

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} p(x|x_i)$$

Suppose that the data objects are partitioned into k^* disjoint clusters, $\{c_j\}_{j=1}^{k^*}$, therefore, $p(x)$ can be written as follows:

$$p(x) = \sum_{j=1}^{k^*} P(c_j) p(x|c_j)$$

where $P(c_j)$ are the probabilities of the clusters and can be estimated by n_j/n, such that n_j is the number of data objects in cluster c_j. The cluster conditional probability distribution $p(x|c_j)$ of X is estimated by

$$p(x|c_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} p(x|x_{i_j})$$

where x_{i_j} is the i^{th} object in cluster c_j. That is, $p(x|c_j)$ is the average of the n_j association distributions $p(x|x_{i_j})$.

But these clusters $\{c_j\}_{j=1}^{k^*}$ and their cluster conditional distributions $p(x|c_j)$ are unknown and represent exactly what we need to find. Therefore, by determining how to group the set of n association distributions, we can compute by aggregation through averaging an empirical cluster conditional probability

distribution of X for each cluster. Section 3 presents an information-theoretic consensus clustering process developed for grouping of association distributions.

After computing $p(x|c_j)$ for all $j \in \{1, \cdots, k^*\}$, a final object assignment (or re-arrangement) is performed where each object x_i is assigned to its most likely cluster c_j which satisfies that $P(X = x_i|c_j) \geq P(X = x_i|c_l)$ for $l \neq j$, and $l \in \{1, \cdots, k^*\}$.

Hence, each object is assigned to a particular cluster $\{c_i\}_{i=1}^{k^*}$. Furthermore, estimated assignment probabilities of objects to a cluster c_j, $P(c_j|x)$ can be computed using Bayes rule as follows.

$$P(c_j|x) = \frac{p(x|c_j)P(c_j)}{p(x)} = \frac{p(x|c_j)P(c_j)}{\sum_{l=1}^{k^*} p(x|c_l)P(c_l)}$$

Notice that the cluster conditional probability distribution introduced here is a discrete probability function defined on the finite space of object labels, rather than the feature vectors as conventionally assumed in supervised classification.

3 The Clustering Combination Process

In this section we present a heuristic information-theoretic method for grouping association distributions into k^* clusters, which as discussed in Section 2.2 is the basis for empirically estimating a cluster conditional probability distribution for each of the combined clusters. We call this method the Jensen-Shannon Divergence based Clustering Combination (JSDCC) as it mainly utilizes the Jensen-Shannon divergence.

We will use the short hand notation $p_i(x)$ and $p_i(x_j)$ to refer to $p(x|x_i)$ and $P(X = x_j|x_i)$ respectively. For details on the information theoretic measures used, the reader is referred to [14]. The Shannon entropy [15] $H(p_i)$ measures the information content of $p_i(x)$ and is given by

$$H(p_i) = -\sum_{j=1}^{n} p_i(x_j) \log p_i(x_j)$$

The Jensen-Shannon JS divergence is a measure of distance between the probability distributions $p_i(x)$, and $p_j(x)$ and is given by

$$JS(p_i, p_j) = H(\bar{p}_{ij}) - \Big(\frac{H(p_i) + H(p_j)}{2}\Big) \tag{1}$$

where $\bar{p}_{ij}(x)$ is the average of the probability distributions $p_i(x)$ and $p_j(x)$.

The JS is symmetric, bounded [16] and $JS(p_i, p_j) = 0 \iff p_i(x) = p_j(x)$, $\forall x$. Furthermore, it can be generalized to measure the divergence between any finite number r of probability distributions, and allows for weighted averages among distributions. The most general form of the JS divergence is given in Equation 2 where $\pi = \{\pi_1, \pi_2, \cdots \pi_r\}$ is the set of normalized weights. In this paper equal weights are used.

$$JS_\pi(\{p_i(x)|1 \le i \le r\}) = H\Big(\sum_{i=1}^{r} \pi_i p_i(x)\Big) - \sum_{i=1}^{r} \pi_i H(p_i) \quad (2)$$

The method starts by selecting k^* prototypes $T = \{t_1, \cdots, t_{k^*}\}$ from the set of association distributions $\{p(x|x_i)\}_{i=1}^{n}$, which will be used as initial representatives of the k^* clusters. From an information theory point of view, we would like to select the most informative prototypes, yet they should be as divergent as possible from each others. Hence, information content measured using the entropy is used to rank the distributions. But, we cannot select the k^* prototypes with the largest entropies, because although they can indeed contain a great deal of information, it may be the same information. So, we need to also choose them so that they are not redundant. Therefore we want to select those prototypes that have maximum divergence among themselves, where divergence is measured using the generalized Jensen-Shannon divergence given in Equation 2. In principle, this involves a search over $\binom{n}{k^*}$ different prototype sets, which is an enormous search for large n. Therefore, we use a incremental method which assesses prototypes individually based on their entropies and their divergences from previously selected prototypes, as described in Algorithm 1.

Algorithm 1 Selection of Prototypes

1: $t_1 \leftarrow \arg\max_{p_i(x)} H(p_i)$
2: **for all** $j \in \{2, \cdots, k^*\}$ **do**
3: **for all** $l \in \{1, \cdots, n\}$ **do**
4: compute Divergence $D(p_l) \leftarrow JS_\pi(p_l, t_l, t_2, \cdots, t_{j-1})$ as given in Equation 2, such that $\pi = 1/j$
5: **end for**
6: select $t_j \leftarrow \arg\max_{p_i(x)} H(\arg\max_{p_i(x)} D(p_i))$
7: **end for**

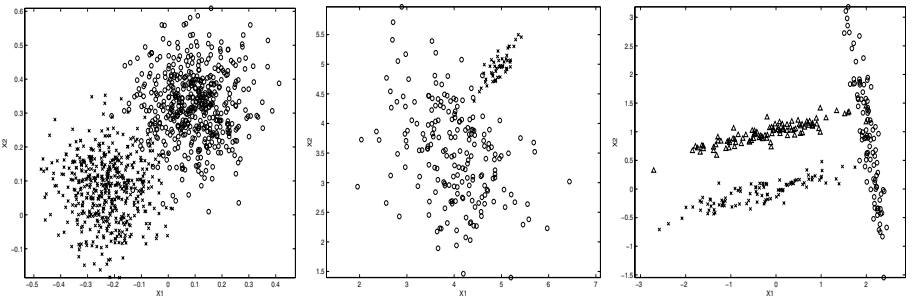
Following the selection of k^* cluster prototypes, a distribution merging procedure is used to merge each of the n association distributions with the prototype with minimum JS divergence. The procedure is summarized in Algorithm 2.

4 Experimental Analysis

Experiments are performed using the K-means clustering algorithm. The number of clusterings generated m are 10 and 50. We use fixed values of $k_i = k$ for all the m clusterings within an ensemble. Different values of k are used starting from the true number of clusters in a dataset and gradually increasing k. The final number of combined clusters k^* represent the number of true clusters and is assumed known. For each combination of k and m, we evaluate the quality of combined clusterings generated using SLink, JSDCC, CSPA and WSnnG using the commonly used F-measure. We also show the average ensemble's F-measure.

Algorithm 2 Merging of distributions

1: **for all** $i \in \{1, \cdots, n\}$ **do**
2: **for all** $j \in \{1, \cdots, k^*\}$ **do**
3: compute $JS(p_i, t_j)$ as given in Equation 1
4: **end for**
5: Let $t \leftarrow \arg\min_{t_j \in T} JS(p_i(x), t_j)$ and merge $p_i(x)$ with t.
6: **end for**
7: Average merged k^* clusters of distributions to estimate $\{p(x|c_i)\}_{i=1}^{k^*}$
8: Assign objects to their most likely clusters and compute assignment probabilities (See Section 2.2)



(a) (b) (c)

Fig. 1. Data sets (a) 2D2K dataset consists of 2 Gaussian clusters, not linearly separable, (b) 2D2C-Non-Spherical dataset has 2 unbalanced non spherical clusters, and (c) 2D3C-Strings has 3 string-like clusters of the same size

We use five dataset to evaluate the performance of the different methods in a number of different situations. The first is the 2D2K dataset used in [3] and downloaded from http://www.strehl.com/. We use a random sample of 300 data points. The dataset is shown in Figure 1 (a) and consists of two Gaussian clusters which are not linearly separable. The second dataset is 2D2C-Non-Spherical shown in Figure 1 (b), which is artificially generated and consists of 2 ellipsoidal clusters with different sizes (200,50), and covariances. The third dataset is 2D3C-Strings, shown in Figure 1 (c), and consisting of three ellipsoidal and elongated clusters of equal sizes (100,100,100), and different orientations. The fourth is the Iris dataset available from the UCI machine learning repository and consists of 4-dimensional points in 3 clusters, one linearly separable and 2 interleaving. The fifth is the Wisconsin diagnostic breast cancer data (WDBC), also available from the UCI repository, and consists of 569 instances and 30 numeric attributes, with class distribution of 357 benign, 212 malignant

Notice that the ensemble approach in [5] used varying k_i. In [3, 8, 9] various clustering techniques were used to generate the ensembles, and in [9], the vote threshold and the number of nearest neighbors were varied. Here, the focus is on evaluating their respective underlying combination methods on the ensembles

Table 1. F-Measure for the 2D2K data set

m	k	Ensemble's Mean	k^*	Consensus Functions SLink	JSDCC	CSPA	WSnnG
10	2	0.986	2	0.986	0.986	0.963	0.963
10	4	0.731	2	0.627	0.986	0.980	0.976
10	6	0.579	2	0.666	0.983	0.973	0.973
10	8	0.535	2	0.666	0.979	0.973	0.976
10	10	0.497	2	0.976	0.983	0.960	0.966
50	2	0.986	2	0.986	0.986	0.963	0.963
50	4	0.725	2	0.704	0.983	0.980	0.980
50	6	0.582	2	0.665	0.979	0.980	0.980
50	8	0.533	2	0.665	0.983	0.976	0.976
50	10	0.480	2	0.665	0.979	0.976	0.976
Average				0.7606	0.9827	0.9724	0.9729

Table 2. F-Measure for the 2D2C-Non-Spherical

m	k	Ensemble's Mean	k^*	Consensus Functions SLink	JSDCC	CSPA	WSnnG
10	2	0.743	2	0.769	0.769	0.729	0.725
10	4	0.643	2	0.984	0.976	0.718	0.722
10	6	0.546	2	0.984	0.984	0.729	0.722
10	8	0.471	2	0.725	0.924	0.722	0.725
10	10	0.414	2	0.729	0.608	0.722	0.725
50	2	0.716	2	0.769	0.769	0.729	0.733
50	4	0.637	2	0.984	0.972	0.729	0.729
50	6	0.544	2	0.984	0.984	0.722	0.726
50	8	0.474	2	0.984	0.980	0.733	0.729
50	10	0.415	2	0.984	0.984	0.726	0.729
Average				0.8896	0.8950	0.7259	0.7265

specified by the parameters m and k as shown in the first two columns of the results Tables 1, 2, 3, 4, and 5.

From the results, it is noted that in cases of 2D2K, 2D3C-Strings, Iris, and WDBC there is a decline observed in terms of the F-measure with the SLink approach which is believed to be due to its inherent chaining effect particularly when the clusters are not linearly separable. The CSPA and WSnnG did not adapt successfully to the cluster structure of the 2D2C-NonSperical, whereas the SLink and JSDCC performed well in most ensembles, by uncovering the cluster structure which in this case the K-means (see average at $k = 2$) has failed to find. The 2D3C-Strings was the hardest on the K-means (see average at $k = 3$), whereas at some combination of m and k the performance of JSDCC, CSPA and WSnnG was relatively good in discovering the cluster structure, yet a dependency on k is clearly observed. In the case of the Iris, JSDCC, CSPA and WSnnG outperformed SLink, and in case of WDBC, JSDCC was best.

Table 3. F-Measure for the 2D3C-Strings Dataset

		Ensemble's		Consensus Functions			
m	k	Mean	k^*	SLink	JSDCC	CSPA	WSnnG
10	3	0.652	3	0.663	0.690	0.7	0.695
10	5	0.631	3	0.752	0.700	0.686	0.722
10	7	0.644	3	0.664	0.669	0.729	0.669
10	9	0.561	3	0.590	0.926	0.959	0.873
10	11	0.513	3	0.576	0.811	0.929	0.896
50	3	0.655	3	0.690	0.690	0.659	0.635
50	5	0.628	3	0.725	0.690	0.635	0.722
50	7	0.627	3	0.751	0.751	0.641	0.736
50	9	0.568	3	0.699	0.924	0.926	0.809
50	11	0.511	3	0.699	0.618	0.953	0.923
Average				0.6809	0.7469	0.7817	0.7680

Table 4. F-Measure for the Iris dataset

		Ensemble's		Consensus Functions			
m	k	Mean	k^*	SLink	JSDCC	CSPA	WSnnG
10	3	0.872	3	0.891	0.891	0.853	0.839
10	5	0.759	3	0.758	0.831	0.966	0.946
10	7	0.708	3	0.758	0.890	0.96	0.919
10	9	0.637	3	0.758	0.831	0.78	0.826
10	11	0.590	3	0.758	0.831	0.80	0.753
10	13	0.520	3	0.771	0.973	0.973	0.886
50	3	0.846	3	0.891	0.891	0.853	0.839
50	5	0.755	3	0.831	0.897	0.919	0.893
50	7	0.684	3	0.758	0.831	0.973	0.946
50	9	0.622	3	0.758	0.966	0.979	0.693
50	11	0.573	3	0.758	0.897	0.979	0.973
50	13	0.525	3	0.758	0.973	0.973	0.686
Average				0.7873	0.8918	0.9173	0.8499

Table 5. F-Measure for the WDBC Data

		Ensemble's		Consensus Functions			
m	k	Mean	k^*	SLink	JSDCC	CSPA	WSnnG
10	2	0.844	2	0.844	0.844	0.675	0.635
10	3	0.799	2	0.890	0.890	0.818	0.802
10	4	0.764	2	0.682	0.779	0.839	0.792
10	5	0.679	2	0.683	0.855	0.694	0.692
10	6	0.594	2	0.683	0.877	0.840	0.844
50	2	0.844	2	0.844	0.844	0.675	0.635
50	3	0.799	2	0.743	0.890	0.815	0.801
50	4	0.764	2	0.682	0.821	0.797	0.792
50	5	0.678	2	0.683	0.855	0.700	0.692
50	6	0.588	2	0.683	0.854	0.844	0.675
Average				0.7417	0.8509	0.7697	0.7360

The average performance over all the ensemble configurations shown for each dataset is as follows. In the case of the 2D2K dataset, JSDCC improves over the lowest performing combination method by 29%. In the case of the 2D2C-Non-Spherical, JSDCC was best and improves over the lowest by 23% and by 20% over the K-means at the true number of cluster (k=2). In the case of 2D3C-Strings, it is lower by 5% than the higest performing method and improves over the lowest by 8% and by 14% over the K-means at true number of clusters (k=3). In the case of the Iris data, it is 2% lower than the highest and improving by 14% over the lowest. Finally, in the case of WDBC, it is the best, and improves over the lowest combination method by 16%.

5 Discussion and Conclusion

The co-association matrix represents a voting medium allowing the generation of consensus clustering. In this paper, we proposed a probabilistic model based on the co-association matrix in which the definition of association distributions allowed for the generation of empirical cluster conditional association distributions for each of the combined clusters. The model allowed the evaluation of the probabilities with which objects belong to the each cluster. The key objectives of developing this model are to represent the associations of neighboring data objects and to allow for the manipulation of their association distributions using probabilistic and information theoretic tools. Noticeably, the model expresses of the same idea of shared nearest neighbors [17, 8, 9], since the association distribution is indeed a function of each object's co-associated neighbors. A Jensen-Shannon divergence based Clustering Combination (JSDCC) method was developed for grouping of association distributions.

The JSDCC method has a quadratic computational complexity ($O(n^2)$), since the computation of the JS divergence is $O(n)$. Future work will focus on optimizing the approach for scalability to large datasets. For instance, we can cut down the number of distance computations by processing all the objects that has high s_{ij} values with selected prototypes. That is, we can use both the s_{ij} values and divergences jointly, leaving divergence computation for less obvious cases. Although the space complexity of the co-association matrix itself is $O(n^2)$, it is possible in future work to limit the number of co-associated objects to the K-nearest neighbors which will reduce the space requirement to $O(Kn)$.

Acknowledgements

We would like to thank the anonymous referees for their helpful comments. This work was partially funded by an NSERC strategic grant.

References

1. A.K. Jain, M.N Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.
2. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

3. A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining partitionings. In *Conference on Artificial Intelligence (AAAI 2002)*, pages 93–98, Edmonton, July 2002. AAAI/MIT Press.
4. A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, 3:583–617, December 2002.
5. A. Fred and A.K. Jain. Data clustering using evidence accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition. ICPR 2002*, volume 4, pages 276–280, Quebec City, Quebec, Canada, August 2002.
6. A. Fred and A. K. Jain. Robust data clustering. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2003*, Madison - Wisconsin, USA, June 2003.
7. Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik. Voting-merging: An ensemble method for clustering. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks-ICANN 2001*, pages 217–224, Vienna, Austria, August 2001. Springer.
8. H. Ayad and M. Kamel. Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. In *Multiple Classifier Systems: Fourth International Workshop, MCS 2003, UK, Proceedings.*, pages 166–175, 2003.
9. H. Ayad and M. Kamel. Refined shared nearest neighbors graph for combining multiple data clusterings. In *The 5th International Symposium on Intelligent Data Analysis IDA 2003. Berlin, Germany, Proceedings. LNCS. Springer.*, 2003.
10. A. Topchy, A.K. Jain, and W. Punch. Combining multiple weak clusterings. In *IEEE Intl. Conf. on Data Mining 2003, Proceedings*, pages 331–338, Melbourne, Fl., November 2003.
11. George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. Technical Report TR 95-035, Department of Computer Science and Engineering, University of Minnesota, 1995.
12. B. Fischer and J.M. Buhmann. Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):513–518, 2003.
13. S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.
14. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, USA, 1991.
15. C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.
16. J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1995.
17. R.A. Jarvis and E.A. Patrick. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers*, C-22(11):1025–1034, November 1973.