 CMES, vol.97, no.2, pp.175-197, 2014

Enrichment Procedures for Soft Clusters: A Statistical Test and its Applications

R.D. Phillips[1], M.S. Hossain[1], L.T. Watson[1,2], R.H. Wynne[3], and Naren Ramakrishnan[1]

Abstract: Clusters, typically mined by modeling locality of attribute spaces, are often evaluated for their ability to demonstrate 'enrichment' of categorical features. A cluster enrichment procedure evaluates the membership of a cluster for significant representation in predefined categories of interest. While classical enrichment procedures assume a hard clustering definition, this paper introduces a new statistical test that computes enrichments for soft clusters. Application of the new test to several scientific datasets is given.

Keywords: Cluster enrichment, fuzzy clustering, statistical significance test.

1 Introduction

Clustering is an unsupervised process that models locality of data samples in attribute space to identify groupings: samples within a group are closer to each other than to samples from other groups. To assess whether the discovered clusters are meaningful, a typical procedure is to see if the groupings capture other categorical information *not originally used during clustering*. For instance, in microarray bioinformatics, data samples correspond to genes and their expression vectors, clusters capture locality in expression space, and they are evaluated to see if genes within a cluster share common biological function/annotations. (Observe that the functional annotations are not used during clustering). In text mining, data samples correspond to documents and their text vectors, clusters capture locality in term space, and are evaluated for their correspondence with *a priori* domain information such as topics. In remote sensing, data samples correspond to pixels in an image,

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0106, USA.

[2] Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123, USA.

[3] Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, USA.

clusters capture locality of pixel intensities, and are evaluated for their correspondence with land cover classifications.

All of the above applications are essentially determining whether locality in one space preserves correspondence with information in another space, also referred to as the *cluster assumption* [Chapelle, Schölkopf, and Zien (2006)]. While cluster evaluation is typically conducted as a distinct post-processing stage after mining, recently developed clustering formulations blur this boundary. For instance, in Wagstaff, Cardie, Rogers, and Schrödl (2001), locality information is used along with background knowledge to influence the clustering. Such background knowledge takes the form of constraints, some of which dictate that certain samples should appear in the same cluster, while others specify that two samples should be in different clusters. Similarly, in Tishby, Pereira, and Bialek (1999), clusters are designed using an objective function that balances compression of the primary random variable against preservation of mutual information with an auxiliary variable. With the advent of semisupervised clustering [Chapelle, Schölkopf, and Zien (2006)], more ways to integrate labeled and unlabeled information are rapidly being proposed.

The design of both classical and the newer clustering algorithms is predicated on the ability to evaluate clusters for enrichment and using this information to drive the refinement and subsequent discovery of clusters. However, classical statistical enrichment procedures (e.g., using the hyper-geometric distribution [Ewens and Grant (2001)]) assume a hard clustering formulation. The focus here is on soft clusters where the groupings are defined by portions of individual samples. This paper presents a new statistical test to enrich soft clusters and demonstrates its application to several datasets.

2 Clustering

Clustering can be used to analyze and discover relationships in large datasets. Strictly unsupervised clustering is used in the absence of information about target clusters and variables of interest, however, clustering can be partially supervised or guided when additional information regarding target clusters is available.

Clustering by itself does not correspond to classification, the process by which class labels are assigned to individual data elements, but clustering can be a useful tool in the classification of large datasets. When clusters are used to organize similar elements in a dataset, class labels can be assigned to entire clusters, allowing individual elements within a cluster to be assigned that class label. Because samples or elements in a particular cluster are similar or "close," they are assumed likely to share a class label, which is known as the *cluster assumption*. Assigning labels

to a modest number of clusters is less time intensive than assigning labels to many individual samples, so if the cluster assumption holds, clustering is an efficient and powerful tool in classification. Unfortunately, this cluster assumption does not hold in all cases, as there is no rule that dictates that "close" samples must share a label.

Finally, the descriptions of clustering, semisupervised clustering, and cluster evaluation given above assume a specific type of clustering where clusters are collections of individual elements, known as hard or crisp clustering. Alternatively, clusters can be defined by portions of individual samples, known as soft clustering. Soft cluster evaluation becomes less intuitive as clusters will no longer "contain" individual samples, and clusters cannot be composed primarily from samples belonging to one class in the same sense. The following subsections define hard and soft clustering and classification.

2.1 Hard Clustering

Hard clustering produces clusters that are a collection of individual samples. Let the ith sample be denoted by $x^{(i)} \in \Re^b$ where $i = 1, \ldots, n$. A cluster is typically represented by a prototype, such as the mean of the samples contained in the cluster, and let the jth cluster prototype be $U^{(j)} \in \Re^b$ where $j = 1, \ldots, K$. All clusters taken together form a partition of the data, defined by a partition matrix, w with $w_{ij} = 1$ indicating that the ith sample belongs to the jth cluster, $w_{ij} = 0$ otherwise, and $\sum_{j=1}^{K} w_{ij} = 1$ for all i. Each sample is a member of exactly one cluster.

A classic example of a simple hard clustering method is the K-means clustering algorithm that locates a local minimum point of the objective function

$$J(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \rho_{ij} \tag{1}$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1, \qquad \text{for } i = 1, \ldots, n,$$

where $\rho_{ij} = \|x^{(i)} - U^{(j)}\|_2^2$ [MacQueen (1974)]. In this case, ρ_{ij} is a measure of dissimilarity or distance between the ith sample and the jth cluster. The K-means clustering algorithm attempts to find the ideal partition that minimizes the sum of squared distances between each sample and the prototype of the cluster to which the sample belongs. The algorithm for K-means requires K initial cluster prototypes and iteratively assigns each sample to the closest cluster using

$$w_{ij} = \begin{cases} 1, & \text{if } j = \underset{1 \le j \le K}{\operatorname{argmin}} \rho_{ij}, \\ 0, & \text{otherwise,} \end{cases}$$

for each i, followed by the cluster prototype (mean) recalculation

$$U^{(j)} = \sum_{i=1}^{n} (w_{ij} x^{(i)}) \Big/ \sum_{i=1}^{n} w_{ij}$$

once w has been calculated. This process, guaranteed to terminate in a finite number of iterations, continues until no further improvement is possible, terminating at a local minimum point of (1).

In hard clusters, such as those produced by K-means, the collection of samples that belong to a particular cluster can be evaluated to determine a cluster's eligibility to perform classification. The class memberships of the labeled samples in a particular cluster can be modeled using discrete random variables generated from binomial, multinomial, or hypergeometric distributions, for example. These random variables form the basis of statistical tests used to evaluate clusters for classification. For example, let V_{ic} be a Bernoulli random variable where success ($V_{ic} = 1$) indicates the ith labeled sample is labeled with the cth class. The number of labeled samples labeled with the cth class in a particular cluster would be a binomial random variable $V_{c,j} = \sum_{i \in I_j} V_{ic}$ where I_j is the index set of labeled samples belonging to the jth cluster. This binomial random variable can be used as the basis for a statistical hypothesis test to determine if the number of samples labeled with the cth class (as opposed to all other classes) in the jth cluster is significant. In practice, the cth class that would be tested would be the class that is most represented in the jth cluster, or mathematically, $c = \text{argmax}_{1 \le c \le C} V_{c,j}$ for a particular j where C is the number of classes.

2.2 Soft Clustering

Soft clusters are clusters that instead of containing a collection of individual samples, contain portions of individual samples. Another view of soft clustering is that each sample has a probability of belonging to a particular cluster. Soft clustering has advantages over hard clustering in that a sample is not simply assigned to the closest cluster, but information is preserved about relationships to other clusters as well. Furthermore, these continuous assignments are less constrained that discrete assignments, resulting in a less constrained objective function. Like in hard clustering, w_{ij} indicates cluster membership, but instead of being either zero or one, $w_{ij} \in (0,1)$, and like in hard clustering, $\sum_{j=1}^{K} w_{ij} = 1$ for all i. Some versions of fuzzy clustering do not impose this requirement, but those nonprobabilistic methods will not be considered here.

An example of a soft clustering method analogous to K-means is fuzzy K-means

that locates a local minimum point of the objective function

$$J(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^{p} \rho_{ij} \tag{2}$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1$$

where ρ_{ij} is still the squared Euclidean distance between $x^{(i)}$ and $U^{(j)}$ and $p > 1$ [Bezdek (1980)]. The algorithm that minimizes this objective function is similar to that of K-means in that it first calculates

$$w_{ij} = \frac{(1/\rho_{ij})^{1/(p-1)}}{\sum_{k=1}^{K} (1/\rho_{ik})^{1/p-1}}$$

for all i and j followed by calculating updated cluster prototypes

$$U^{(j)} = \sum_{i=1}^{n} w_{ij}^{p} x^{(i)} \Big/ \sum_{i=1}^{n} w_{ij}^{p}.$$

The cluster prototype is a weighted average. This iteration (recalculation of the weights followed by recalculation of cluster prototypes, following by recalculation of the weights, etc.) is guaranteed to converge (with these definitions of ρ_{ij}, $U^{(j)}$, and w_{ij}) for $p > 1$ [Bezdek (1980)].

3 Soft Cluster Evaluation

Evaluation of soft clusters requires taking cluster weights into account when examining class memberships of the labeled samples. Each labeled sample will have some positive membership in each cluster, and a new type of evaluation will be necessary to directly evaluate soft clusters. Soft cluster memberships could be converted to hard cluster memberships for the purpose of cluster evaluation, but if soft clustering is warranted, those soft clusters should be evaluated directly.

Hard cluster evaluation (for classification) is based on the composition of the cluster, or what type of samples are making up the cluster. The question of whether a cluster should be used for classification can be answered when some of the samples within the cluster have labels and there are a sufficient number of samples to draw statistical conclusions. Because soft clusters no longer "contain" samples,

the more important question is whether the relative magnitudes of memberships between samples of a particular class and the cluster are significantly different. In other words, if the magnitude of cluster memberships for samples of a particular class appear to be significantly higher than memberships for other classes, then the cluster is demonstrating characteristics of that class. With hard clusters, a cluster is pure if only one class is contained in the cluster; no samples labeled with another class are present in the cluster. This is impossible in soft clustering as all types of samples will have positive memberships in all clusters, and in practice, these memberships, although possibly small, will be nonnegligible.

Just as hard clusters that are ideal for classification contain only one class, soft clusters that are ideal for classification will be representative of just one class. The goal in using soft clustering for classification is to assign a class label to an entire cluster (the same goal for hard clusters), but just as each sample has a soft membership in a particular cluster, each sample will have soft membership in a class. The samples demonstrate characteristics of multiple classes, justifying soft classification, but the clusters (logical grouping of similar data) should not contain or represent multiple classes. The goal of this work is to associate a soft cluster to one particular class if that class is clearly dominant within the cluster. Probability will determine how clearly a particular cluster is composed of one class, and if this probability passes a predetermined threshold test, the cluster will be associated with a class.

3.1 Hypothesis Test

The statistical tests used to evaluate clusters in this paper are statistical hypothesis tests, where a null hypothesis is proposed. If observed evidence strongly indicates the null hypothesis should be rejected, the alternate hypothesis will be accepted. In the absence of compelling evidence to the contrary, the null hypothesis cannot be rejected.

The first hypothesis test is based on the average cluster weights in the cluster of interest, the jth cluster. In order to associate the jth cluster to the cth class, the average cluster weight for the cth class

$$\bar{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij},$$

where n_c is the number of samples labeled with the cth class and J_c is the index set of samples labeled with the cth class, should be statistically significantly higher than other cluster weights for the jth cluster. If the weights for samples labeled with the cth class are higher in general than samples from arbitrary classes, the cluster is demonstrating a tendency to the cth class, and can be used to discriminate the cth class from other classes.

The null hypothesis is that the average cluster weights for samples from the cth class in the jth cluster is not significantly different from the average cluster weight for samples from all classes in the jth cluster. The alternate hypothesis is that the average weight for samples from the cth class in the jth cluster is significantly different (higher) than the average cluster weight for all samples. Note that in practice, only the class with the highest average cluster weight for the jth cluster would be considered. Suppose that a test statistic derived for this test is normally distributed, and is in fact a standard normal random variable Z. Then if the observed value is $\hat{z}$, if $P(Z \geq \hat{z}) \leq \alpha$ for $0 < \alpha < 1$, the null hypothesis is rejected. The following sections derive appropriate test statistics to use in this hypothesis test.

3.2 Test Statistic 1

Suppose a dataset x contains n samples $x^{(i)} \in \Re^B$, $i = 1, \ldots, n$. For K fixed cluster centers $U^{(k)} \in \Re^B$, $k = 1, \ldots, K$, the assigned weight of the ith pixel to the jth cluster is

$$w_{ij} = \frac{1/\|x^{(i)} - U^{(j)}\|_2^2}{\sum_{k=1}^{K} 1/\|x^{(i)} - U^{(k)}\|_2^2}$$

which is the inverse of the distance squared over the sum of the inverse squared distances. (Such inverse distance weights are widely used, e.g., by Shepard's algorithm for sparse data interpolation.) Note this is the specific case in the soft clustering algorithm described above when $p = 2$. In many practical applications where a dataset is to be clustered (such as the clustering of a remotely sensed image), it is reasonable to assume that $x^{(i)}$, $i = 1, \ldots, n$ are generated from a finite number of multivariate normal distributions. The act of clustering assumes that the data are generated from a finite number of distributions. The following theorem from Phillips, Watson, Wynne, and Ramakrishnan (2009a) demonstrates that under these assumptions (samples are generated from a finite number of normal distributions), the Lindeberg condition is satisfied and therefore the central limit theorem applies to the sum of a sequence of cluster weight random variables $\sum_{i=1}^{n} W_{ij}$. Let $q = \psi(i)$ denote the distribution from which the random vector $X^{(i)}$ was sampled.

Theorem: Let $X^{(i)}$, $i = 1, 2, \ldots$, be B-dimensional random vectors having one of Q distinct multivariate normal distributions. For $i = 1, 2, \ldots$ and $j = 1, \ldots, K$ define the random variables

$$W_{ij} = W_j(X^{(i)}) = \frac{1/\|X^{(i)} - U^{(j)}\|_2^2}{\sum_{k=1}^{K} 1/\|X^{(i)} - U^{(k)}\|_2^2}$$

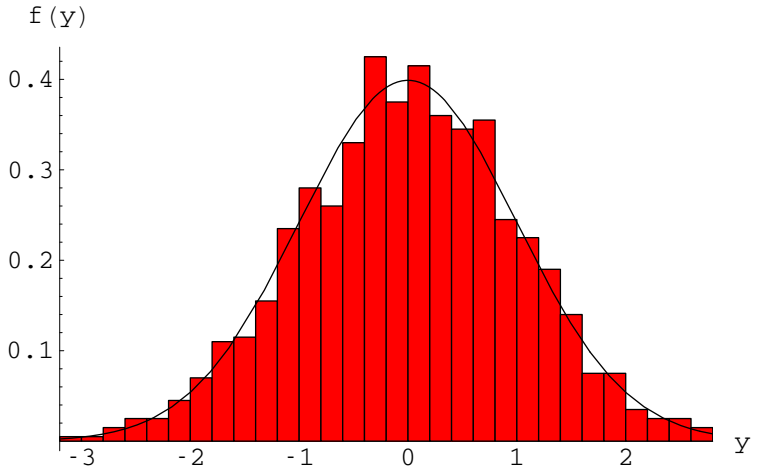


Figure 1: Distribution of sums of weights in one soft cluster out of two.

where K is the number of clusters and $U^{(k)} \in \Re^B$ is the kth cluster center (and is considered fixed for weight calculation). Then for any $j = 1, \ldots, K$,

$$P\left\{\frac{1}{B_{nj}}\sum_{i=1}^{n}(W_{ij} - a_{ij}) < x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz$$

as $n \to \infty$, where $a_{ij} = \mathrm{E}[W_{ij}]$, $b_{ij}^2 = \mathrm{Var}[W_{ij}]$, and $B_{nj}^2 = \sum_{i=1}^{n} b_{ij}^2$.

Remark: The assumption that the $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of normal distributions is stronger than necessary. The proof in Phillips, Watson, Wynne, and Ramakrishnan (2009a) holds if $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of arbitrary distributions.

Experimental clustering results using a dataset described in Section 4 of this paper match this theoretical result, as illustrated by one experiment in Fig. 1. This illustration shows the distribution of sums of cluster weights for one particular cluster (when $K = 2$).

Starting with the normal approximation for the sum of the cluster weights, the standard normal test statistic would be

$$\hat{z} = \frac{\sum_{i \in J_c}(w_{ij} - E[W_{ij}])}{\sqrt{\sum_{i \in J_c}\mathrm{Var}[W_{ij}]}},$$

where $\mathrm{E}[W_{ij}]$ is the expected value of W_{ij} and $\mathrm{Var}[W_{ij}]$ is the variance of W_{ij} for the jth cluster. $\mathrm{E}[W_{ij}]$ and $\mathrm{Var}[W_{ij}]$ are unknown, but can be reasonably approximated using the sample mean

$$\bar{w}_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij}$$

and sample standard deviation

$$S_{\bar{w}_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(w_{ij}-\bar{w}_j)^2}.$$

The Wald statistic is then

$$\hat{z} = \frac{\sqrt{n_c}(\bar{w}_{c,j}-\bar{w}_j)}{S_{\bar{w}_j}},$$

where

$$\bar{w}_{c,j} = \frac{1}{n_c}\sum_{i\in J_c} w_{ij}.$$

Since $\hat{z}$ is generated (approximately) by the standard normal distribution, this test statistic can be used in the proposed hypothesis test.

3.3 Test Statistic 2

One potential issue with the above statistic is that the sample mean and standard deviation calculations assume the sample is identically distributed, which is specifically *not* the assumption in this case (clustering assumes that the data are generated from a number of distributions, where the true number of clusters is equal to the number of distributions, which is unknown apriori). A better statistic acknowledges that the data are not identically distributed, but are generated from a finite number of distributions. Since the number of distributions and the distributions are unknown, the number of classes and the individual class labels, which are assumed to correspond to inherent structure of the data, are used to approximate the true mean and variance of multiple clusters. Precisely, assume that all labeled sample indices i with distribution index $\psi(i) = q$ correspond to the same class label $\phi(i) = c$. If $i \in \psi^{-1}(q)$, then $i \in \phi^{-1}(c)$, but $i \in \phi^{-1}(c)$ does not imply $i \in \psi^{-1}(q)$ (more than one distribution can correspond to one class), and $J_c = \phi^{-1}(c) = \{i \mid \phi(i) = c, 1 \le i \le n\}$. The above statistic requires modification to use class information. In the previous statistic,

$$\sum_{i\in J_c} w_{ij} = \sum_{i=1}^{n} w_{ij}\delta_{phi(i),c},$$

$$\hat{z} = \frac{\sum_{i=1}^{n} (w_{ij}\delta_{phi(i),c} - E[W_{ij}\delta_{\phi(i),c}]}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}[W_{ij}\delta_{\phi(i),c}}},$$

$$\begin{aligned}
&\sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - E[W_{ij}\delta_{\phi(i),c}]) \\
&\quad = \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - a_{ij}\delta_{\phi(i),c}) \\
&\quad = \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - a_{qj}\delta_{\phi(i),c}),
\end{aligned}$$

recalling that $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ for $i \in I_q$. Assume when $\phi(i) = c$, and distribution index $q = \psi(i)$ corresponds to $c = \phi(i)$, then α_{qj} can be approximated by γ_{cj}, the mean of class $c = \phi(i)$. Ideally α_{qj} should be approximated directly, but there is no way to know $\psi^{-1}(q)$, so essentially $\psi^{-1}(q) \subset \phi^{-1}(c)$ is being approximated by $\phi^{-1}(c)$. Unfortunately, using the sample mean of the cth class and the jth cluster to approximate γ_{cj} and therefore α_{qj} breaks down because the sample mean of the cth class and the jth cluster is both the random variable on the left side and the approximation of the expected value on the right side of the minus sign. This is illustrated below. Approximating γ_{cj} (and α_{qj}) with the sample mean for the cth class,

$$\gamma_{cj} \approx \bar{w}_{c,j} \frac{\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n} \delta_{\phi(k),c}}$$

the numerator of the test statistic $\hat{z}$ becomes

$$\begin{aligned}
&\sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - \bar{w}_{c,j}\delta_{\phi(i),c}) \\
&\quad = \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \frac{\sum_{i=1}^{n} w_{kj}\delta_{\phi(k),c}}{\sum_{k=1}^{n} \delta_{\phi(k),c}} \sum_{i=1}^{n} \delta_{\phi(i),c} \\
&\quad = \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c} = 0.
\end{aligned}$$

Thus this test statistic does not work because the value being tested is the same as the estimated mean for the cth class.

In order to make use of class information to estimate distribution statistics (mean and variance), it is necessary to modify the random variable to model class labels as well as cluster memberships. Consider each labeled sample's membership in a particular class, say the cth class, to be a Bernoulli trial V_{ic}, where $V_{ic} = 1$ indicates the ith sample is labeled with the cth class, and W_{ij} is defined above. Define

$$Y_{c,j} = V_{1c}W_{1j} + V_{2c}W_{2j} + \cdots + V_{nc}W_{nj},$$

where n is the total number of labeled samples as the random variable for the sum of weights for samples in the cth class to the jth cluster. The Central Limit Theorem applies to this sum of bounded random variables with finite mean and variance (see Theorem 1), and $Y_{c,j}$ is approximately normal.

Consider now the test statistic

$$\hat{z} = \frac{y_{c,j} - E[Y_{c,j}]}{\sqrt{\text{Var}[Y_{c,j}]}}.$$

Fixing j and c, assuming W_{ij} and V_{ic} are independent, and defining $m_q = |I_q|$, the number of indices i for which $X^{(i)}$ has the qth distribution,

$$\begin{aligned} E[Y_{c,j}] &= E\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n} E[W_{ij}V_{ic}] \\ &= \sum_{i=1}^{n} E[W_{ij}]E[V_{ic}] = \sum_{q=1}^{Q} m_q \alpha_{qj} p_c = p_c \sum_{q=1}^{Q} m_q \alpha_{qj}, \end{aligned}$$

where p_c is the probability that $V_{ic} = 1$. Assuming all the samples are independent and recalling that $\text{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$ where $i \in I_q$,

$$\begin{aligned} \text{Var}[Y_{c,j}] &= \text{Var}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n} \text{Var}[W_{ij}V_{ic}] \\ &= \sum_{i=1}^{n} (E[W_{ij}^2 V_{ic}^2] - E[W_{ij}V_{ic}]^2) \\ &= \sum_{i=1}^{n} (p_c E[W_{ij}^2] - p_c^2 a_{ij}^2) \\ &= \sum_{i=1}^{n} (p_c (b_{ij}^2 + a_{ij}^2) - p_c^2 a_{ij}^2) \\ &= \sum_{q=1}^{Q} m_q (p_c (\beta_{qj}^2 + \alpha_{qj}^2) - p_c^2 \alpha_{qj}^2) \\ &= p_c \sum_{q=1}^{Q} m_q (\beta_{qj}^2 + (1 - p_c)\alpha_{qj}^2). \end{aligned}$$

In the above formula, p_c would be approximated by its maximum likelihood estimate $n_c/n = |J_c|/n$. In order to estimate α_{qj}, assume that the qth distribution corresponds to the cth class, $\psi^{-1}(q) \subset \phi^{-1}(c)$, and

$$\alpha_{qj} \approx \bar{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}, \quad c = 1, \ldots, C,$$

where C is the number of classes. Then

$$\begin{aligned} E[Y_{c,j}] &= p_c \sum_{q=1}^{Q} m_q \alpha_{qj} \approx p_c \sum_{d=1}^{C} n_d \cdot \frac{1}{n_d} \sum_{i \in J_d} w_{ij} \\ &= \frac{n_c}{n} \sum_{i=1}^{n} w_{ij} = n_c \bar{w}_j, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} m_q (\beta_{qj}^2 + (1-p_c)\alpha_{qj}^2) \\ &\approx p_c \sum_{d=1}^{C} n_d (S_{\bar{w}_{d,j}}^2 + (1-p_c)\bar{w}_{d,j}^2), \end{aligned}$$

where

$$S_{\bar{w}_{d,j}}^2 = \frac{1}{n_d - 1} \sum_{i \in J_d} (w_{ij} - \bar{w}_{d,j})^2$$

Using these expressions for the mean and variance of $Y_{c,j}$, the Wald statistic for the cth class and jth cluster is

$$\hat{z} = \frac{y_{c,j} - n_c \bar{w}_j}{\sqrt{p_c \sum_{d=1}^{C} n_d (S_{\bar{w}_{d,j}}^2 + (1-p_c)\bar{w}_{d,j}^2)}}$$

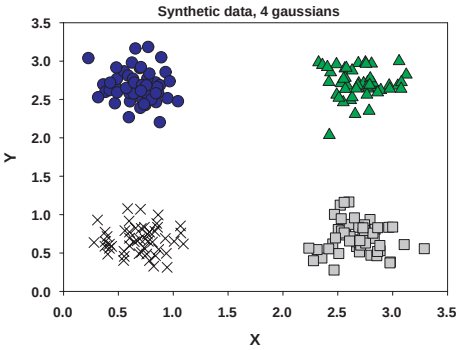
and the null hypothesis is rejected if $P(Z \geq \hat{z}) \leq \alpha$.

4 Experimental Results

This section presents experimental results to demonstrate the functioning of the new statistical test. It is important to distinguish the nature of enrichments identified by the new test from the quality of clusters mined by a specific algorithm. The features evaluated are (i) whether the test is able to recognize clusters with partial

Table 1: Datasets

Dataset	# of instances	# of features	# of classes
Synthetic	200	2	4
Ionosphere	351	34	2
Vehicle	846	18	4
Glass	214	9	6
Cardiotocography	2126	21	10
Breast Tissue	106	9	6
Steel Plates Faults	1941	27	7



(a) Synthetic data (four Gaussians).

2.53E-33	3.84E-05	7.33E-05	7.49E-05
6.97E-05	7.04E-05	2.28E-33	3.92E-05
8.67E-05	8.51E-05	4.25E-05	2.66E-33
3.70E-05	2.75E-33	6.95E-05	7.76E-05

(b) Wald statistic (soft assignments).

1.73E-34	4.46E-05	4.46E-05	4.46E-05
4.46E-05	4.46E-05	1.73E-34	4.46E-05
4.46E-05	4.46E-05	4.46E-05	1.73E-34
4.46E-05	1.73E-34	4.46E-05	4.46E-05

(c) Wald statistic (hard assignments).

2.20E-48	1	1	1
1	1	2.20E-48	1
1	1	1	2.20E-48
1	2.20E-48	1	1

(d) Enrichment with hypergeometric distribution.

Figure 2: Enrichments of synthetic data (Jaccard similarity between class labels and clusters=1.0).

memberships (soft assignments) as being significant, (ii) whether it leads to a higher number of assignments in soft clustering situations, and (iii) the variation in number of enrichments as entropy of clusters and significance levels are changed. For the purpose of this evaluation, consider the soft k-means algorithm where membership probabilities at each stage of the iteration are non-zero across the clusters.

Table 1 describes the datasets used in this study; with the exception of the synthetic dataset, all are taken from the UCI KDD/ML data repository. In each case, the number of clusters to be identified is set equal to the number of natural classes present in the dataset.

Fig. 2 presents results on synthetic data involving four separable Gaussians in a two-dimensional layout. The enrichment p-values are also shown for all 16 combinations for the soft and hard versions of the Wald statistic as well as the hypergeometric test, which is commonly used for cluster evaluation. As can be seen, the qualitative trends are the same so that for all stringent thresholds the results yield four clusters enriched with four different class labels.

0.0002	0.0031
0.0007	0.0059

0.0038	0.0225
0.2341	0.5811

6.66E-15	1
1	6.66E-15

Figure 3: Ionosphere data. *(left)* Soft assignments: p-values are derived from Wald statistic for the cth class and jth cluster. *(middle)* Hard assignments: p-values are derived from Wald statistic for the cth class and jth cluster. *(right)* Enrichment with hypergeometric distribution. (Jaccard similarity between fuzzy k-means and the actual class-labels: 0.5865.)

Fig. 3 presents a more complicated situation with the ionosphere dataset. This dataset involves two classes and there are more tangible differences between the three statistical tests. Note that the Jaccard similarity between the fuzzy k-means and class labels is not a perfect 1. As a result, for various values of the p-value threshold, it is possible to get one, two, three, or four cells enriched by the Wald statistic (soft assignment) whereas the hypergeometric distribution can lead to only two or four cells enriched. The Wald statistic (hard assignment) also performs better than the hypergeometric distribution.

Fig. 4 more directly describes a plot of the number of enriched cells as the p-value cutoff is varied, using the vehicle dataset. The Wald statistics lead to a consistently greater number of enrichments compared to the hypergeometric test. A similar plot can be seen in Fig. 5.

A different type of evaluation is shown in Fig. 6(a) where the membership probabilities are artificially varied (from a hard membership) to impose a specified entropy

6.44E-20	4.11E-13	1.97E-11	0.004603
1.40E-05	0.010169	0.6351	9.90E-09
0.024838	3.03E-05	0.000149	7.72E-05
0.025948	5.46E-08	0.001541	4.93E-06

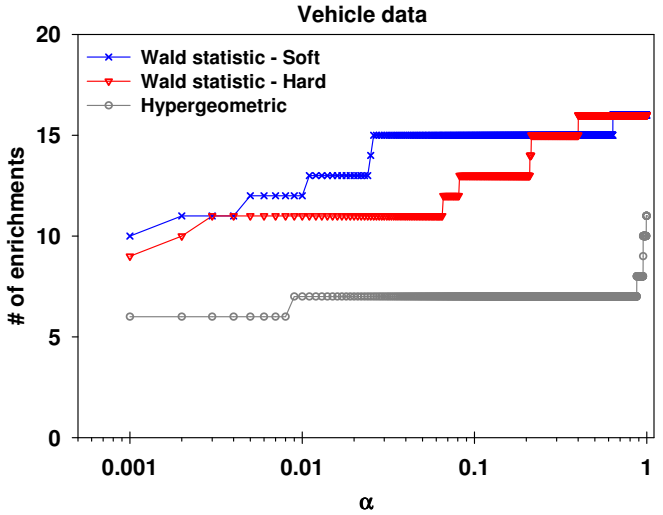
(a) Wald statistic (soft assignments).

4.19E-22	1.03E-11	1.30E-08	0.081319
1.07E-11	0.065437	0.40054	2.64E-11
0.21376	0.000141	0.001119	7.38E-05
0.20959	3.72E-06	0.002103	9.07E-06

(b) Wald statistic (hard assignments).

1.16E-28	1	1	0.008583
1	0.99319	0.87421	1.35E-21
0.95129	2.57E-06	8.34E-05	1
0.95245	1.78E-08	0.000202	1

(c) Enrichment with hypergeometric distribution.



(d) Number of enrichments at different p-value cut-offs.

Figure 4: Vehicle data. (a) Soft assignments: p-values are derived from Wald statistic for the cth class and jth cluster. (b) Hard assignments: p-values are derived from Wald statistic for the cth class and jth cluster. (c) Enrichment with hypergeometric distribution. (Jaccard similarity between fuzzy k-means and the actual class-labels: 0.6506.) (d) Number of enrichments at different p-value cut-offs with the three enrichment procedures.

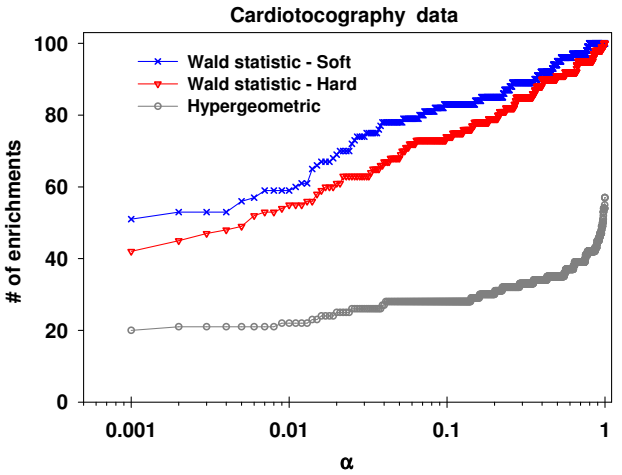


Figure 5: Cardiotocography data. Number of enrichments at different p-value cut-offs with the three enrichment procedures (Jaccard similarity between fuzzy k-means and actual class labels: 0.7896).

on their distribution. As the entropy increases, the number of enrichments drops monotonically in the case of the Wald (soft) statistic whereas the hypergeometric enrichment test does not account for the entropy in a smooth manner. Fig. 6(b) demonstrates the variation for a fixed value of the entropy but increasingly lax values of the p-value threshold. Again, the enrichments for the Wald (soft) statistic increase steadily. Similar plots for the breast tissue, steel plate faults, and glass datasets are shown in Figs. 7, 8, 9, respectively. Finally, Fig. 10 superimposes the variation of p-value cutoff and entropy threshold to describe how the variation seen in previous plots manifests at all p-value thresholds, whereas the hypergeometric distribution is uniformly unable to provide a richer variety of enrichments.

5 Conclusion

This paper presented a new statistical test suitable for enrichment of soft clusters. It was shown how this test produces significantly more enrichments, tunable control of number of enrichments, and smoother variation in enriched cells with entropy and p-value cutoffs. The method can be used as given here or embedded inside a cluster refinment algorithm for continuous evaluation and updating of clusters. Since few soft cluster enrichment methods exist, the framework here contributes a key methodology for clustering and cluster evaluation research.

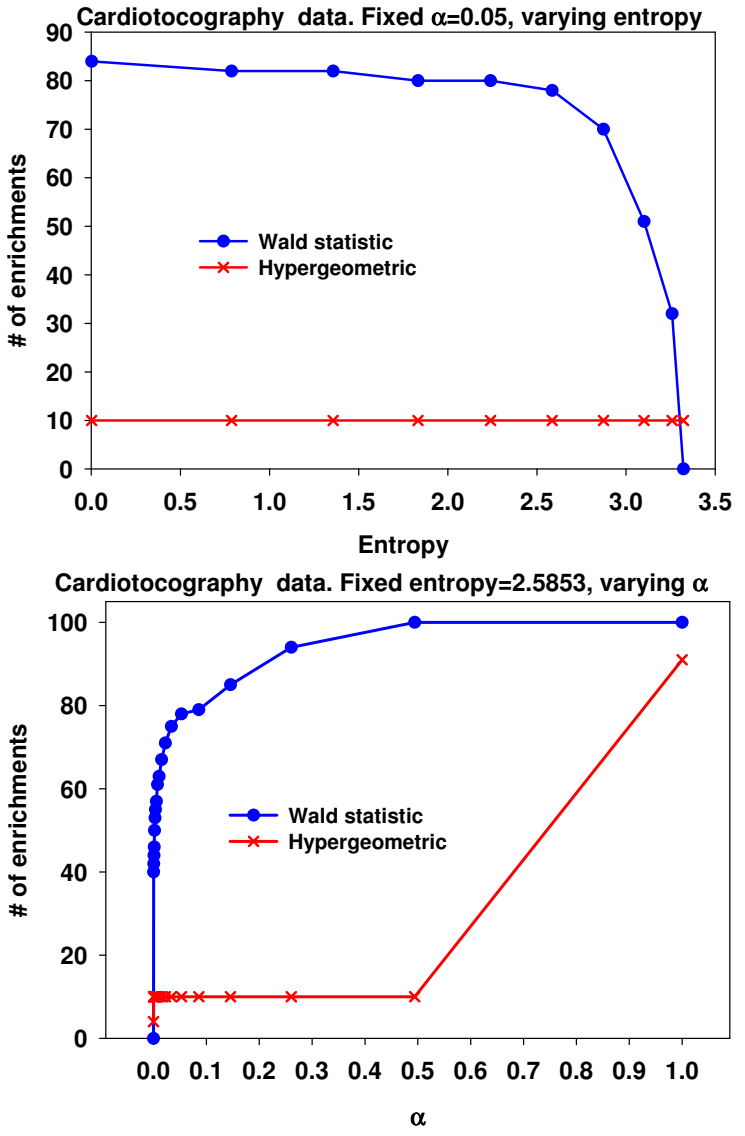
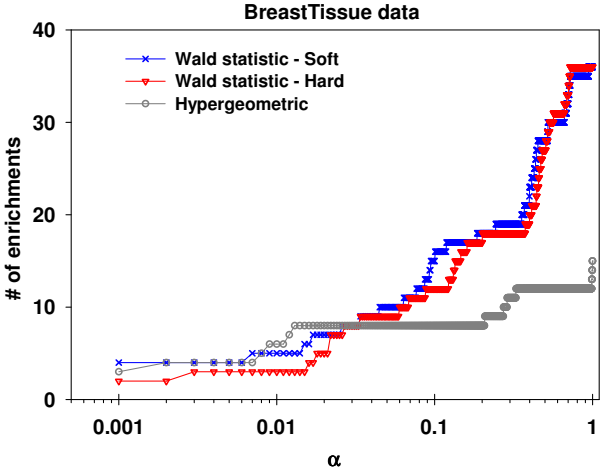
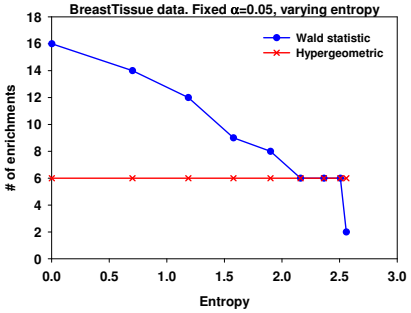


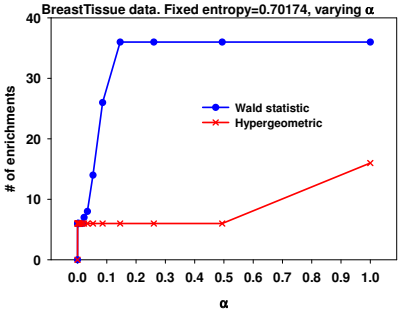
Figure 6: Cardiotocography data. (top) Number of enrichments with fixed p-value threshold but varying entropy. Note that the number of enrichments falls monotonically with increasing entropy. (bottom) Number of enrichments with fixed entropy and varying p-value threshold. Note that the number of enrichments monotonically increases with increasing p-value threshold.



(a) Number of enrichments at different p-value cut-offs.

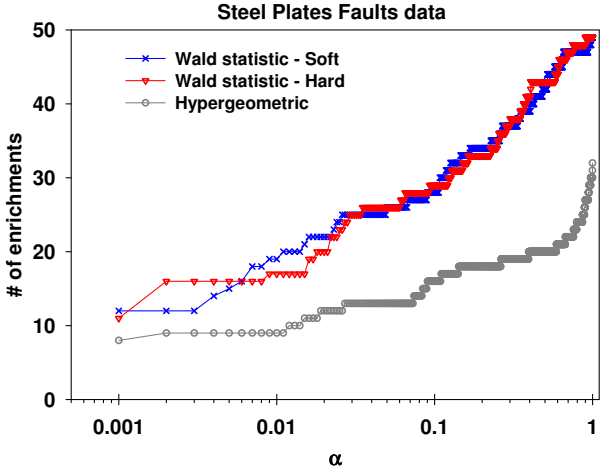


(b) Number of enrichments with fixed p-value threshold and varying entropy.

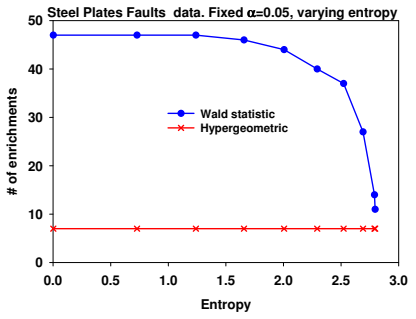


(c) Number of enrichments with fixed entropy threshold and varying p-value threshold.

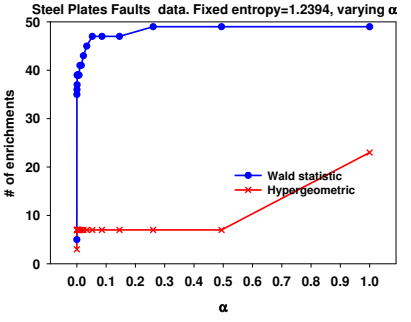
Figure 7: Breast tissue data. (Jaccard similarity between fuzzy k-means and the actual class-labels: 0.7051.) (a) Number of enrichments at different p-value cut-offs with the three enrichment procedures. (b) Number of enrichments with fixed p-value threshold but varying entropy. Note that the number of enrichments falls monotonically with increasing entropy. (c) Number of enrichments with fixed entropy and varying p-value threshold. Note that the number of enrichments monotonically increases with increasing p-value threshold.



(a) Number of enrichments at different p-value cut-offs.



(b) Number of enrichments with fixed p-value threshold and varying entropy.



(c) Number of enrichments with fixed entropy and varying p-value threshold.

Figure 8: Steel plates faults data. (Jaccard similarity between fuzzy k-means and the actual class-labels: 0.6681.) (a) Number of enrichments at different p-value cut-offs with the three enrichment procedures. (b) Number of enrichments with fixed p-value threshold but varying entropy. Note that the number of enrichments falls monotonically with increasing entropy for the Wald statistic. (c) Number of enrichments with fixed entropy and varying p-value threshold. Note that the number of enrichments monotonically increases with increasing p-value threshold.

0.001818	0.52675	0.024238	0.000976	0.002223	0.000158
0.50349	0.001906	0.027655	0.44946	0.01467	0.000578
0.74637	0.11826	0.27557	0.13109	0.14334	0.10098
0.076722	0.077435	0.0341	5.72E-14	0.52862	0.96004
0.22361	0.28552	0.92325	0.001095	0.59563	0.50917
0.001022	0.002282	0.48559	0.90422	0.071899	2.46E-29

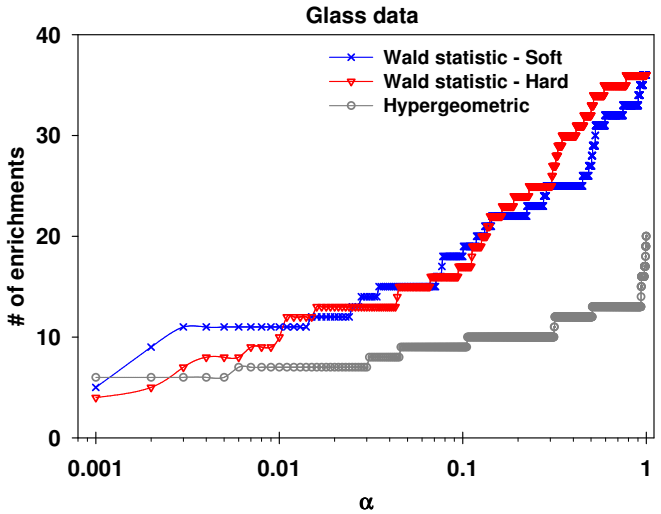
(a) Wald statistic (soft assignments).

0.001963	0.11255	0.12676	0.009157	0.002023	0.000818
0.30894	0.00686	0.003817	0.30555	0.010584	0.000413
0.32133	0.015633	0.45692	0.2294	0.50024	0.13556
0.043498	0.16394	0.5157	1.58E-17	0.14028	0.77611
0.094326	0.11177	0.58895	0.010253	0.044467	0.41467
0.002171	0.066556	0.32999	0.34745	0.18909	6.81E-26

(b) Wald statistic (hard assignments).

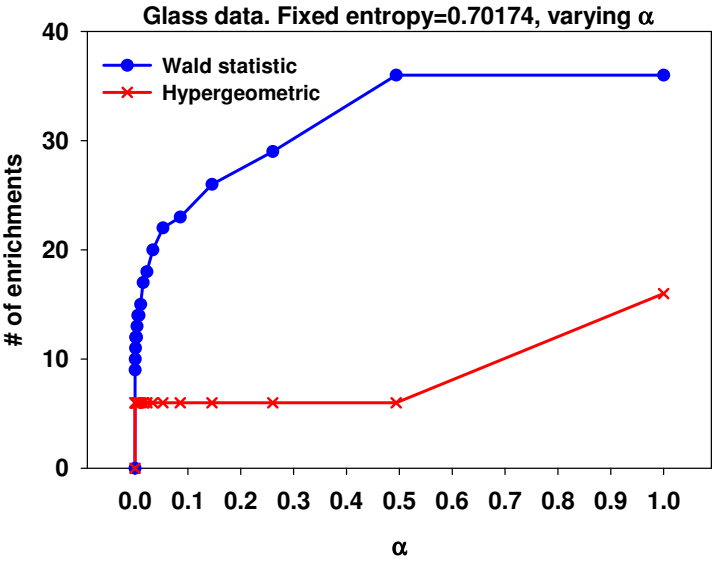
2.94E-05	0.9909	1	1	0.000116	1
0.10504	0.000219	0.000592	0.93647	0.99994	1
0.94019	0.005541	1	1	0.31445	1
1	0.98804	1	2.50E-10	1	0.50722
1	1	1	0.030639	0.045746	0.31695
1	0.99672	1	0.93545	0.9751	1.63E-21

(c) Enrichment with hypergeometric distribution.

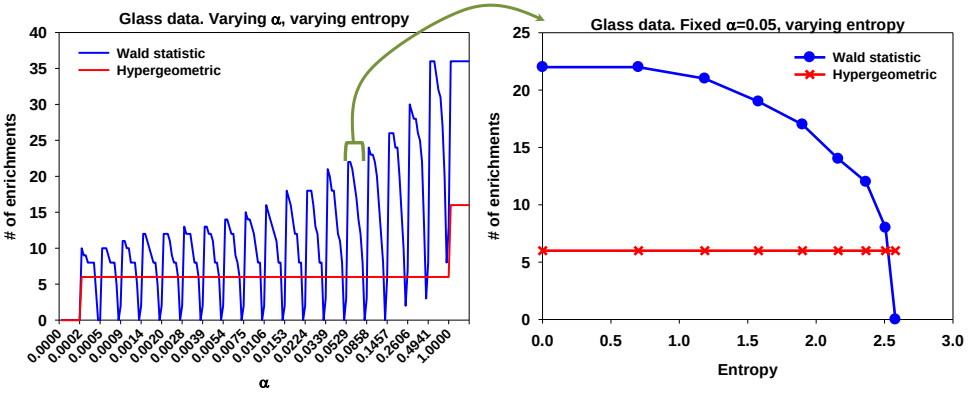


(d) Number of enrichments at different p-value cut-offs.

Figure 9: Glass data. (a) Soft assignments: p-values are derived from Wald statistic for the cth class and jth cluster. (b) Hard assignments: p-values are derived from Wald statistic for the cth class and jth cluster. (c) Enrichment with hypergeometric distribution. (Jaccard similarity between fuzzy k-means and the actual class-labels: 0.7117.) (d) Number of enrichments at different p-value cut-offs with different enrichment procedures.



(a)



(b)

Figure 10: Glass data. In this example, assignments are directly taken from the class labels. The entropy is changed by modifying the membership probability of the class of every instance. (a) Number of enrichments with different p-value thresholds and fixed entropy. (b) The plot at left shows how the number of enrichments change over the p-value thresholds and entropy. Note that the p-value is fixed for each of the spikes in this plot. For example, α remains 0.0020 in the interval between 0.0020 and 0.0028. The plot at the right side shows the change in number of enrichments with entropy where the p-value threshold is fixed.

Acknowledgement: This work was supported in part by Department of Energy Grant DE-FG02-06ER25720 and NIGMS/NIH Grant 5-R01-GM078989.

References

Bezdek, J. (1974): *Fuzzy mathematics in pattern classification*. Ph.D. thesis, Cornell University, Ithaca, NY, 1974.

Bezdek, J. C. (1980): A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 1–8.

Chapelle, O.; Schölkopf, B.; Zien, A. (2006): *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Derraz, F.; Peyrodie, L.; Pinti, A.; Taleb-Ahmed, A.; Chikh, A.; Hautecoeur, P. (2010): Semi-automatic segmentation of multiple sclerosis lesion based active contours model and variational dirichlet process. *Computer Modeling in Engineering & Sciences*, vol. 67, no. 2, pp. 95–118.

Ewens, W.; Grant, G. (2001): *Statistical Methods in Bioinformatics*. Springer.

Gnedenko, B. (1997): *Theory of Probability*. Gordan and Breach Science Publishers, The Netherlands, sixth edition.

Lin, Z.; Cheng, C. (2010): Creative design of multi-layer web frame structure using modified ahp and modified triz clustering method. *Computer Modeling in Engineering & Sciences*, vol. 68, no. 1, pp. 25–54.

MacQueen, J. B. (1967): Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. L.M. Le Cam and J. Neyman, editors, University of California Press.

MacQueen, J. B. (1974): *Fuzzy Mathematics in Pattern Classification*. Ph.D. thesis, 1974.

Musy, R. F.; Wynne, R. H.; Blinn, C. E.; Scrivani, J. A.; Mcroberts, R. E. (2006): Automated forest area estimation via iterative guided spectral class rejection. *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 8, pp. 949–960.

Phillips, R. D.; Watson, L. T.; Wynne, R. H. (2007): Hybrid image classification and parameter selection using a shared memory parallel algorithm. *Comput. Geosci.*, vol. 33, pp. 875–897.

Phillips, R. D.; Watson, L. T.; Wynne, R. H.; Ramakrishnan, N. (2009a): Continuous iterative guided spectral class rejection classification algorithm: Part 1.

Technical report, Department of Computer Science, VPI&SU, Blacksburg, VA, 2009a.

Phillips, R. D.; Watson, L. T.; Wynne, R. H.; Ramakrishnan, N. (2009b): Continuous iterative guided spectral class rejection classification algorithm: Part 2. Technical report, Department of Computer Science, VPI&SU, Blacksburg, VA, 2009b.

Richards, J. A.; Jia, X. (1999): *Remote Sensing Digital Image Analysis*. Springer-Verlag, Berlin, third edition.

Tishby, N.; Pereira, F. C.; Bialek, W. (1999): The information bottleneck method. In *Proc. of 37th annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377.

van Aardt, J. A. N.; Wynne, R. H. (2007): Examining pine spectral separability using hyperspectral data from an airborne sensor: An extension of field-based results. *International Journal of Remote Sensing*, vol. 28, pp. 431–436.

Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. (2001): Constrained k-means clustering with background knowledge. In *ICML '01*, pp. 577–584.