Guided Cluster Discovery with Markov Model

C.H. Li (chli@comp.hkbu.edu.hk) [*]

Department of Computer Science, Hong Kong Baptist University

Abstract. Cluster discovery is an essential part of many data mining applications. While cluster discovery process is mainly unsupervised in nature, it can often be aided by a small amount of labeled data. A probabilistic model on the clustering structure is adopted and a novel unified energy equation for clustering that incorporates both labeled data and unlabeled data is introduced. This formulation is inspired by a force-field model integrating labeling constraint on labeled data and similarity information on unlabeled data for joint estimation. Experimental results show that good clusters can be identified using small amount of labeled data.

Keywords: clustering, semi-supervised learning, Markov model

1. Introduction

In machine learning for classification problems, there are two distinct approaches to learning or classifying data: the supervised learning and un-supervised learning. The supervised learning deals with problem where a set of data are labeled for training and another set of data would be used for testing. The un-supervised learning deals with problem where none of the labels of the data are available. Unsupervised clustering can be broadly classified into whether the clustering algorithm is hierarchical or non-hierarchical. Hierarchical methods often model the data to be clustered in the form of a tree, or a dendrogram [1].

[*] This work is partially support by the Faculty research grant of the Hong Kong Baptist University

The lowest level of the tree is usually each datum as a cluster. A dissimilarity measure is defined for merging clusters at a lower level to form a new cluster at a higher level in the tree. The hierarchical methods are often computationally intensive for large number of samples and is difficult to analyze if there is no logical hierarchical structure in the data.

Non-hierarchical methods divide the samples into a fixed number of groups using some measure of optimality. The most widely used measure is the minimization of the sum of squared distances from each sample to its cluster center. The k-means algorithm, also known as Forgy's method [2] or MacQueen [3] algorithm is a classical algorithm for non-hierarchical unsupervised clustering. However, the k-means algorithm tends to cluster data into even populations and rare abnormal samples in medical problems cannot be properly extracted as individual clusters. Recent progress in clustering includes the modeling of proximity structure [4], the dynamic programming approach to hierarchical clustering using graphs [5] and spectral method to clustering [6]. However, these methods do not make use of prior knowledge on dataset such as possible labels or possible structures within the dataset.

In recent years, important data mining tasks have emerged with enormous volume of data. The labeling of a significant portions of the data for training is either infeasible or impossible. Sufficient labeled data for training are often unavailable in data mining, text categorization and web page classification. A number of approaches have been proposed to combine a set of labeled data with unlabeled data for improving the classification rate. The co-training approach has been proposed to solve the problem of web page classification where the web pages can be represented by two independent representations [7]. The drawback of this co-training approach is that not all data have two independent representations and the algorithm is thus not easy to be generalized. Subsequently, a similar co-training method is invented for combining labeled and unlabeled data by co-training with two learning algorithms [8]. Instead of using two representations of the data, this co-training algorithm uses two learning algorithms. The naive Bayes classifier and the EM algorithm have been combined for classifying text using labeled and unlabeled data [9]. A modified support vector machine and non-convex quadratic optimization approaches have been studied for optimizing semi-supervised learning [10].

In this paper, a novel approach for clustering using guidance is introduced. We model the data as objects in input feature space under the influence of mutual attractive force. The guidance or a priori knowledge will act as anchors and the rest of the data will be attracted towards the different anchors where natural clusters will emerge. In Section Two, the idea of guided cluster discovery is introduced. In Section Three, the force field model and its Markov approximations are introduced. In Section Four, experiments and results will be presented.

2. Guided Cluster Discover and Classification

The guided cluster discovery is closely related to classification problem. In this section, we will look at the similarity and the differences between guided cluster discovery in data mining and the general pattern classification problem.

Suppose the classification task is to classify a set of data denoted by x_i $(i = 1, ..., m)$ into two classes denoted with labels $[A, B]$ respectively. The classification algorithm is to find a corresponding label $y_i \in [A, B]$.

Table I. Classification and Guided Cluster Discovery

	Classification	Guided Cluster Discovery
No. of labeled data $n(x_L)$	Large	Small(a few)
Labeled data	Well-defined	Ill-defined/To be discovered
No. of unlabeled data $n(x_U)$	Small	Large
Use of unlabeled Data x_U	Testing	Training/Analysis

The set of all data in the dataset is denoted as x and the set of all labels y. The cardinality of both x and y is m.

In the classification problem, a fraction of data in the dataset are labeled and the remaining data are to be classified. The set containing all labeled data are denoted as x_L where for each element $x_i \in x_L$, the label y_i for that data is known. Similarly, the set of unlabeled data are denoted by x_U and the set of unknown labels are denoted by y_U. The usual non-intersecting requirement is followed where $x_L \cup x_U = x$, $y_L \cup y_U = y$. Using this notation, the distinction between the data available in data mining and general machine learning is shown in Table I. In traditional classification, a large amount of labeled training data is usually available for training a mapping between the training data and the corresponding labels. In data mining, a large amount of unlabeled data is available and it is often costly or even impossible to assign labels to a significant portion of the unlabeled data for learning.

In order to fully utilize the unlabeled data, guided cluster discovery significantly improves classification accuracy by incorporating unlabeled data for training. In traditional classification, the training phase is carried out by constructing a mapping between the training samples pair though the labeled dataset $\{x_L, y_L\}$. In the estimation phase, the labels of the unknown testing data $x_i \in x_U$ will be obtained. This situation is generally applicable to different areas of pattern recognition and machine learning, especially for situations where the unknown testing data are obtained sequentially in time. However, in the case of data mining, the unlabeled data x_U are already available for the learning algorithm and the only unknown is the set y_U. The guided cluster discovery is similar to semi-supervised learning where there is a small amount of knowledge about the possible labeling in the dataset. When the amount of supervised data is very small, the learning process depends mainly on cluster discovery and a small number of constraints defined by the prior knowledge or assumed hypotheses. As demonstrated by the co-training method in classifying web pages and the color tracking with transductive learning method, the use of unlabeled data x_U can significantly increase the classification accuracy in machine learning tasks.

3. Probability Force Field For Guided Cluster Discovery

The learning of the value of the labels in the dataset is a process that finds the set of labels y_U given the data x_U, x_L and y_L. Instead of a direct estimation on the actual labels y_i we estimate the probability of labels being a given label. For a two class classification problem with class labels $[A, B]$, we represent the probability of the data i taking the labels A as $P_i = P(y_i = A)$. For labeled data $x_i \in x_L$, if $y_i = A$, then $P_i = 1$, else if $y_i = B$, then $P_i = 0$. The probabilistic guided cluster discovery is to estimate the P_i for all i where $x_i \in x_U$.

The construction of the probabilistic force-field model depends on the following assumptions:

– Data vectors with small Euclidean distances between them will have similar probabilities

– Labeled data vectors has fixed probability of either 1 or 0

– The probability P_i of unlabeled data vectors freely distributes themselves to settle in a optimal configurations as defined by energy equation defined by the above two constraint.

The first assumption of spatial close data vector having similar probability can be modeled using attractive force between data vectors in high dimension vector spaces. The force between two data vectors i and j with inverse power law is given by,

$$F_{ij} = \frac{G}{r_{ij}^c} \tag{1}$$

where r_{ij} is the Euclidean distance between i-th vector and j-th vector, c is an integer constant, and G is a fixed positive constant defining the strength of attraction. For example, the gravitation law is an inverse square law with c equals to 2. Suppose that the data vector x_i is in m-dimensional space, we embed the probability into the data vector x_i to form a $m+1$-dimension vector $[x_i, \alpha P_i]$, where α is a constant balancing the scale of the probability and the data vector. The Euclidean distance between two extended vectors in the $m+1$ space is given by,

$$r_{ij} = \sqrt{|x_i - x_j|^2 + \alpha(P_i - P_j)^2} \tag{2}$$

where $|x_i - x_j|$ is the Euclidean distance in the m-dimension space and α is a positive constant balancing the scale of probability to the data vector space.

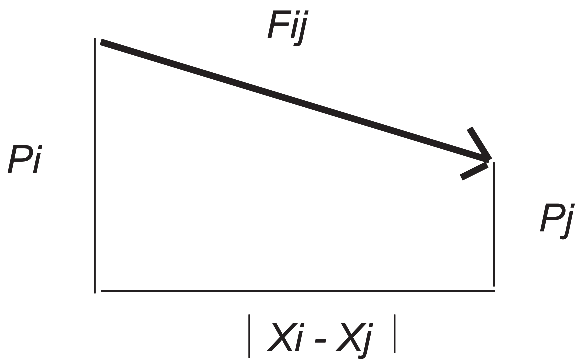

Figure 1. Force vector between two probability vectors

The schematic representation of the embedded vector relationship is shown in Figure 1. The forces between probabilistic vector can then be written as

$$F_{ij} = \frac{G}{\sqrt{|x_i - x_j|^2 + \alpha(P_i - P_j)^2}}. \tag{3}$$

As the data vector x is fixed in spaces, the degree of freedom is along the freely distributable probability. Thus, the effective force on the probability vector is the component of the force along the probability axis

$$Fp_{ij} = \frac{G(P_i - P_j)}{\sqrt{|x_i - x_j|^2 + \alpha(P_i - P_j)^2}}. \tag{4}$$

Alternatively, a force-field energy approach can also be specified. In general, an attractive force can be equivalently represented by a force field energy equation where the energy experienced by a data point i is given by

$$U_i(P) = -\sum_j \frac{G'}{r_{ij}^{c-1}}, \tag{5}$$

where the dependence of P is effected through the dependence on r_{ij}. With P_i and P_j being small, r_{ij} will be minimized and the energy will be lowered. The estimated probability can then be solved by minimizing the energy of the system. The energy of the total system is given by

$$U(P) = -\sum_i \sum_j \frac{G'}{r_{ij}^{c-1}}. \tag{6}$$

An approximation of the above system can be readily obtained by the use of Markov assumptions [11]. Assuming that the interaction is localized by a fixed neighborhood, the above energy equation can be simplified to

$$U(P) = -\sum_i \sum_{j \in N_i} \frac{G'}{r_{ij}^{c-1}}. \tag{7}$$

where N_i is the neighbourhood of the data vector i. There are two choices of neighbourhood in high-dimensional space. First, we can define a hypercube with distance d centered at the data vector i. All data vectors inside this hypercube are elements in N_i. Alternatively, we can define N_i to be the set of k-th nearest neigbours of i. The use of nearest neighbours in constructing the Markov assumptions allows a fixed size neighbourhood for each data i and and a fixed computational $O(kn)$.

4. Results and Discussions

The guided cluster discovery is tested with the iris dataset and two synthetic datasets. Initially, the probability of known labeled samples are assigned to their values and the unknown labels are assigned a random values near 0.5. A gradient descent on the force field with inverse square law is used to update the probability for 1000 iterations. The 8-th nearest neighbourhood system is used in the two following experiments. The only unknown parameter α is specified approximately by balancing the ratio of the range of probability [0,1] to the ratio of the average ranges of data x.

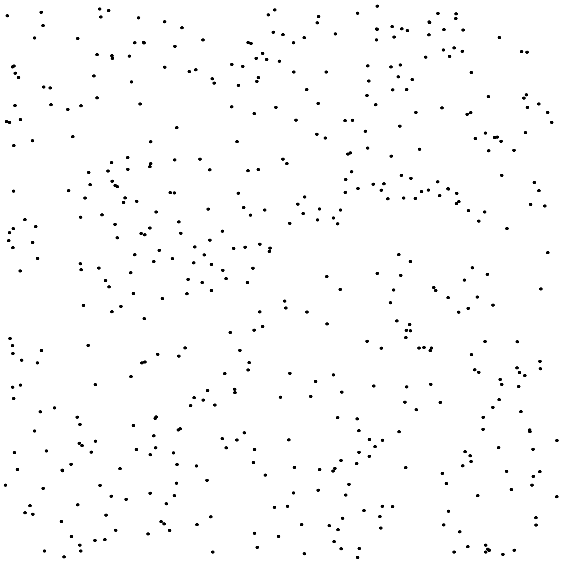

Figure 2. Two clusters

4.1. Cluster Discovery on Non-linear Boundary

A synthetic dataset is used for testing the guided cluster discovery algorithm. Figure 2 shows the data for a two cluster clustering problem. The two clusters are seperated with a sinusoidal boundary. Figure 3 shows the ground truth of the dataset. Results of applying classical cluster discovery algorithms: the k-means algorithm, the hierarchical tree algorithms with single link, average link and complete link are shown in Figure 4, Figure 5, Figure 6, Figure 7 respectively. The clusters discovered by k-means and the average link tree are closer to the ground truth with some errors along the non-linear boundaries. The single link tree and the complete link tree have very poor performance for this dataset.

To test the performance of the guided cluster discovery algorithm on the synthetic data, three training samples are chosen from each classes. Figure 8 shows the training samples where the data with known labels

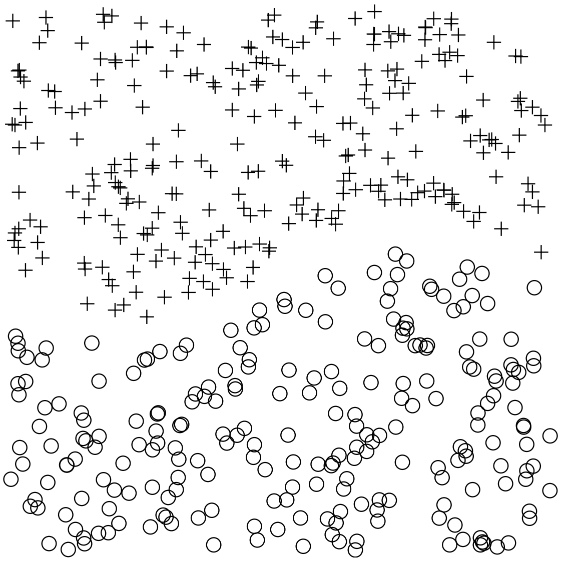

Figure 3. Solution I

are marked with circles and diamond respectively. With a small number of labeled training samples from each class, one can judge from the figures that inference based/decision surface based classifier will not be able to determine accurately the sinusoidal boundary between the cluster. Figure 9 shows the initial probability for the synthetic data. There are 500 data and the first three from each class are selected as training samples. The data with unknown labels are assigned with random probability in the range [0.45 0.55]. Figure 10 shows the initial estimated label, those data with probability above 0.5 is assigned to class I and those data with probability below 0.5 is assigned to class II. This shows that the labels are initially random, except at three pairs of training data.

The first step in force field based guided cluster discovery is the calculation of the 8-th nearest neighbour distance matrix from the data. The scale balancing constant α is chosen as 0.1. After updating the probability for 100 iterations, the probability for the data is shown in

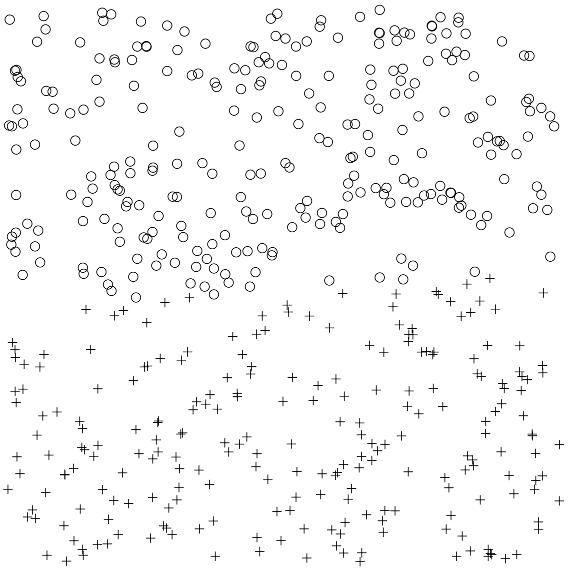

Figure 4. Cluster Discovered by k-means algorithm

Figure 11. The probability is significantly modified after 100 iterations with some of the data point have confidence in having the label value. Those data point with probability close to 0.5 are data point whose label value is not certain. Figure 12 shows the estimated label at 100 iterations, those data with probability above 0.5 is assigned to class I and those data with probability below 0.5 is assigned to class II. There are some errors in the class labels, however those closer to the training samples are mostly correct. The final set of figures show the probability and estimated labels after 1000 iterations. Figure 13 shows that the probability have a wider range where more data are confidently determined after 1000 iterations. Figure 14 shows that the estimated labels after 1000 iterations is very accurate with only 3 errors in the middle left region.

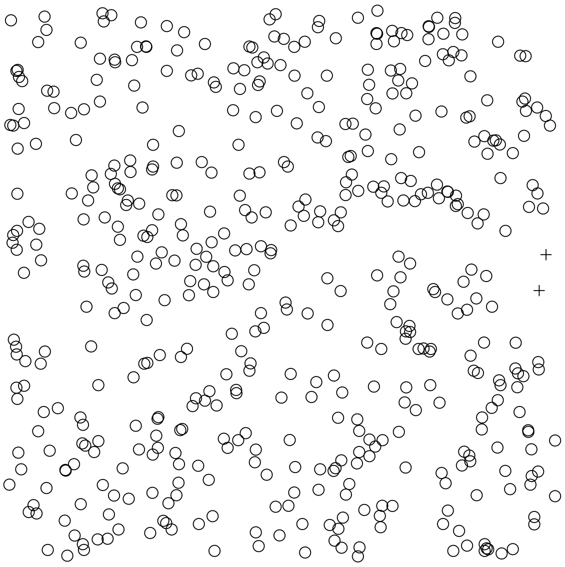

Figure 5. Cluster Discovered by single link tree

4.2. Iris Dataset

The iris dataset is also used for testing the guided cluster discovery. The iris dataset consists of 150 samples of measurement of the iris plant. There are three species of iris in the dataset and each species has 50 samples. Typical approaches uses 100 samples for training and 50 untrained samples for testing. The iris dataset is well studied and results can be found in numerous literature [12].

In the first experiment in guided cluster discovery for iris dataset, we considered using only 2 labeled data. The use of such small amount of training data enables a 'what-if' scenario in data mining to be handled. Even if the true class is unknown, we can choose a few samples randomly and then evaluated the resultant classifications.

The first step in force field based guided cluster discovery is the calculation of the 8-th nearest neighbour distance matrix. After the calculation of the 8-th nearest neighbour distance, we observed that all

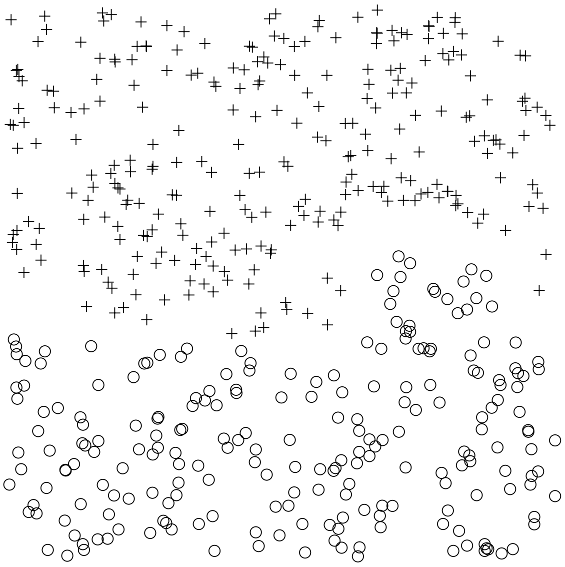

Figure 6. Cluster Discovered by average link tree

the 50 samples from class I are nearest neighbours to members of its own class and thus these samples can be isolated without any further classification. This is also in good agreement with previous work in the iris dataset that the first class is linearly separable and can be classified trivially. Now we concentrate the classification of the second class and the third class. We pick the first sample in class II and first sample in class III as the labeled sample and denote this experiment as experiment (a). The initial probability before update is shown in Figure 15. The first sample is the labeled sample for class II and has a value of one, the 51 sample is the labeled sample for class III and has zero probability being class II. The updated probability distribution is shown in Figure 16. The scale constant α is set as 0.05. The probabilities of data from 1 to 50 belongs to class II, and their values are significantly due to the attractive force among each other and the attractive force from the labeled data. The forces of attraction can propagated through each other and different data feels a different level of attractive force

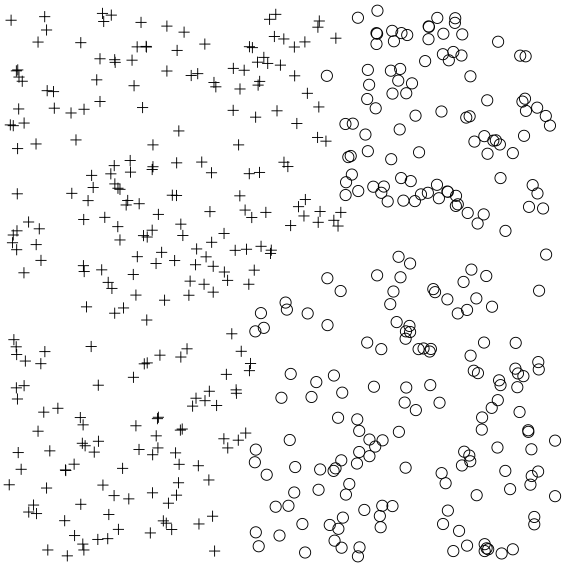

Figure 7. Cluster Discovered by complete link tree

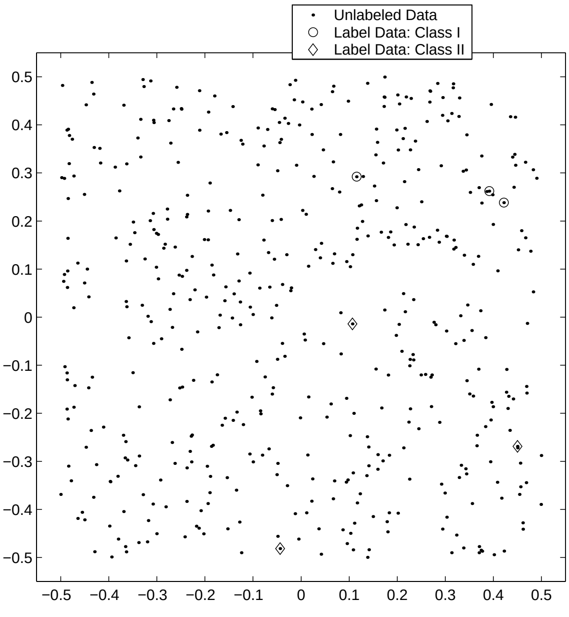

Figure 8. Training Samples

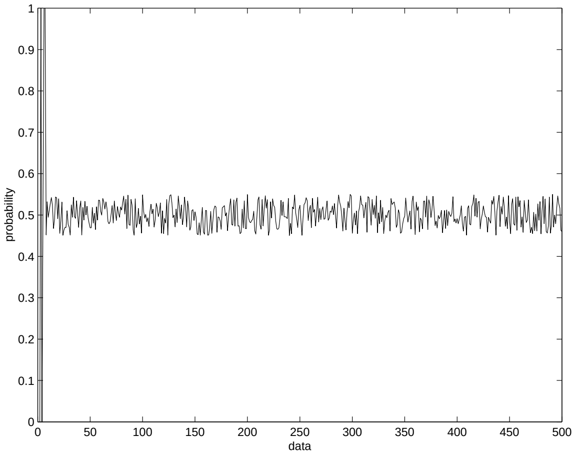

Figure 9. Initial Probability

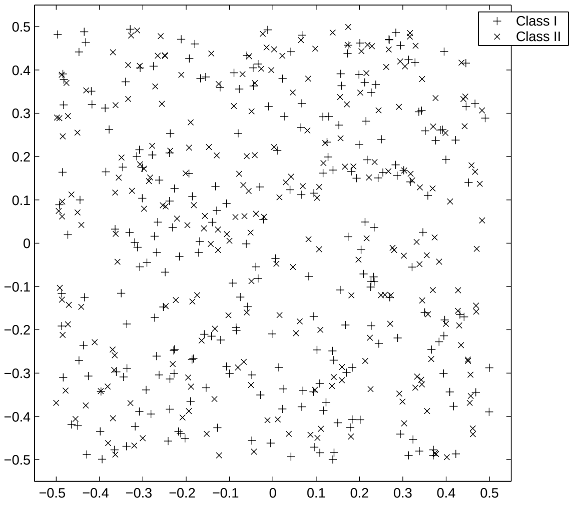

Figure 10. Initial estimated labels

according to their relative positions of the labeled data and other data vectors. The probabilities of data from 51 to 100 belongs to class III and their probabilities are significantly lower than those data that belongs to class II.

In the second experiment, experiment (b), we increase the number of training samples to 2 samples per class. The first two data from class II is used as samples for class II and plant 51 and 52 as samples from

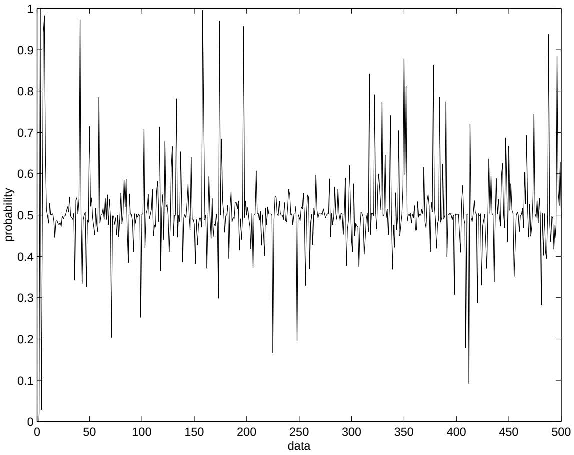

Figure 11. Probability after 100 iterations

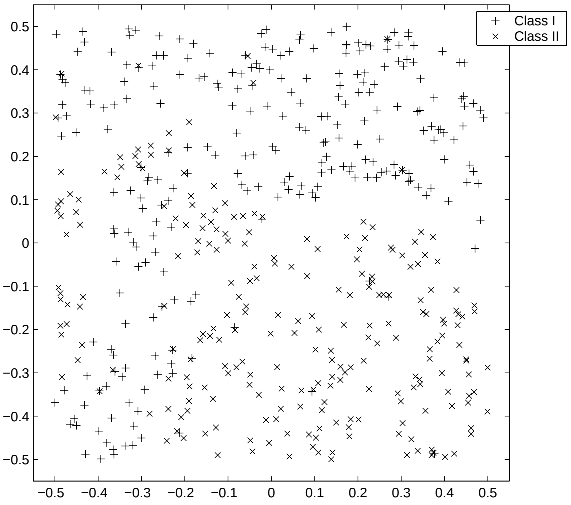

Figure 12. Estimated labels after 100 iterations

class III. The updated probability distribution is shown in Figure 17. The probabilities between the data in class II and class III are much better seperated in this case. For the samples in the first 50 data, only three have probabilities below 0.5. For the samples from 51 to 100, there is only one data with probabilities above 0.5.

Table II shows the number of misclassifications of the two experiments in guided cluster discovery of the iris dataset. The true labels are

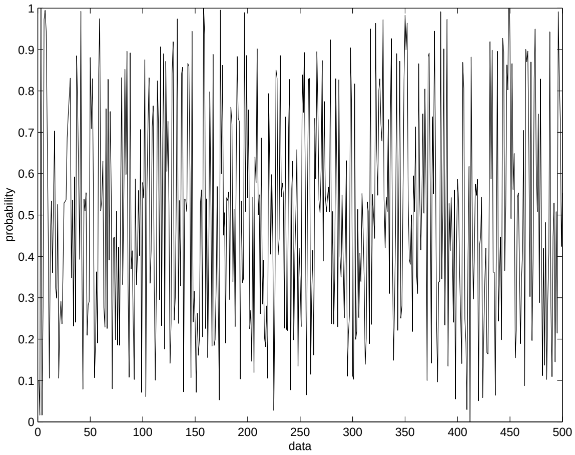

Figure 13. Final Probability

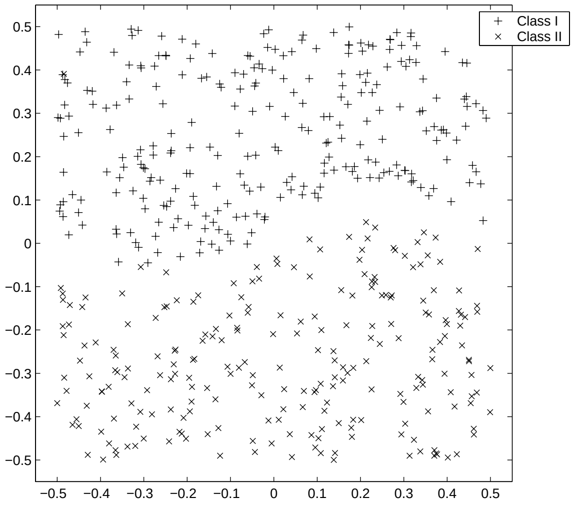

Figure 14. Final estimated labels

taken as the class probabilities above 0.5. Similar results are obtained using other labeled samples as training data. In general, it is found that if the labeled samples represents typical features in the respective class, the learning performances would be acceptable. If a selected sample has feature vector that is similar to samples from the other class, the performances of the learning would be significantly lower.

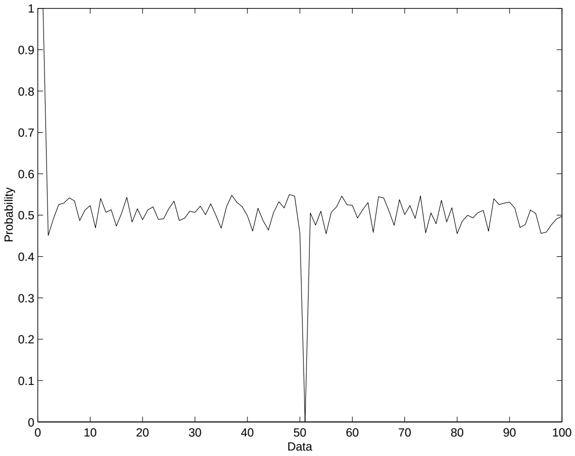

Figure 15. Initial probability of class membership

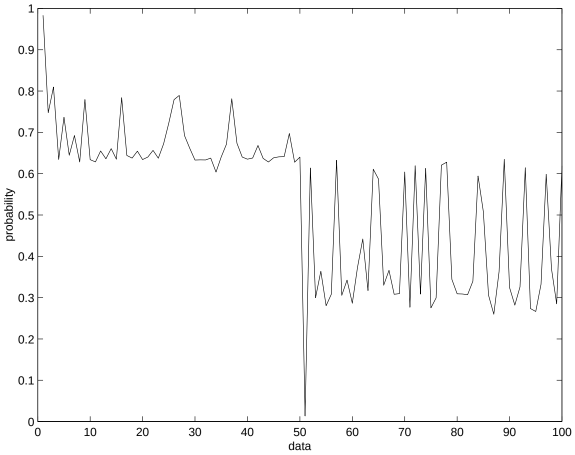

Figure 16. Updated Probability of class membership: Experiment (a)

Table II. Number of errors after guided cluster discovery on iris dataset

	errors in Class II	errors in Class III
Exp. (a)	0	15
Exp. (b)	3	1

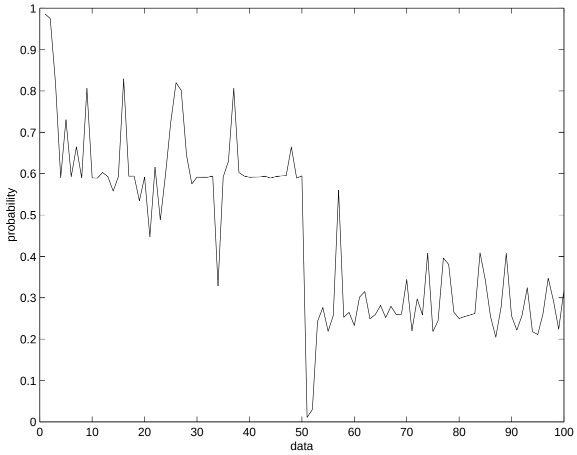

Figure 17. Updated Probability of class membership: Experiment (b)

The errors achieved by the guided cluster discovery algorithm is similar to classical classification algorithm. However, classical classification algorithm often requires up to 100 training samples for achieving this accuracy.

4.3. Gaussian Clusters

The third experiments deals with classification of three Gaussian Clusters. Figure 18 shows the ground truth of the three Gaussian clusters. The three Gaussians Clusters are of different variances and different sizes. To test the force-field method, random points are selected from each cluster as a labeled data from each cluster. Figure 19 shows the training samples used for guided cluster discovery. The initial probabilities estimation is calculated differently from the previous two experiments. In this experiment, we use the nearest neighbour classifier for setting up the initial probabilities for guided cluster discovery. The result of the nearest neighbour classification on the dataset is shown in Figure 20. As there are only one random sample chosen as the labeled sample, the initial classification result has quite a large amount of errors. However, this initial classification result is utilized as the setting up the initial probabilities for transductive learning. The initial probabilities of the i-th data being class j, denoted as p_i^j, is assigned with the following rules:

- assign $p_i^j = 1$ if $y_i = j$ where $y_l \in y_L$,
- assign $p_i^j = 0$ if $y_i \neq j$ where $y_l \in y_L$,
- assign $p_i^j = 0.75$ if x_i is classified by the Nearest neighbour classifier to be class j, and
- assign $p_i^j = 0.25$ if x_i is classified by the Nearest neighbour classifier to be any class other than j.

The initial probabilities for the Class I p^1 is shown in Figure 21. As the data from 1 to 512 belongs to class I, those data with probabilities value 0.25 are data which are incorrectly classified by the nearest-neighbour classifier. The guided cluster discovery algorithm is applied to the probabilities for 1000 iterations with α set as 0.1. Figure 22 shows the probabilities after the iterations. The probabilities of for data from 1 to 512 is raised significantly. The only data with values below 0.5 are located around number 310. From data 513 to 703, most of the data have probabilities below 0.5 and are thus correctly inferred as data not belonging to class I. There are only two data with probabilities above 0.3 in this portion. The estimation procedure can be repeated for estimating the other two class labels. The final estimated labels after calculating all probabilities of the three classes are shown in Figure 23. The estimated labels are in very good agreement with ground truth and visual assessment based on proximity criterion.

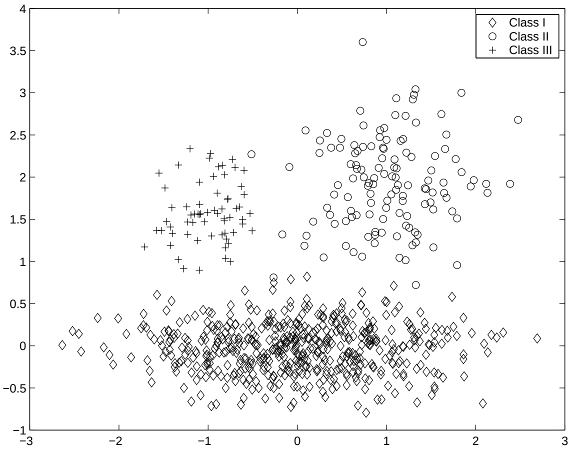

Figure 18. Ground Truth of 3 Gaussian Clusters

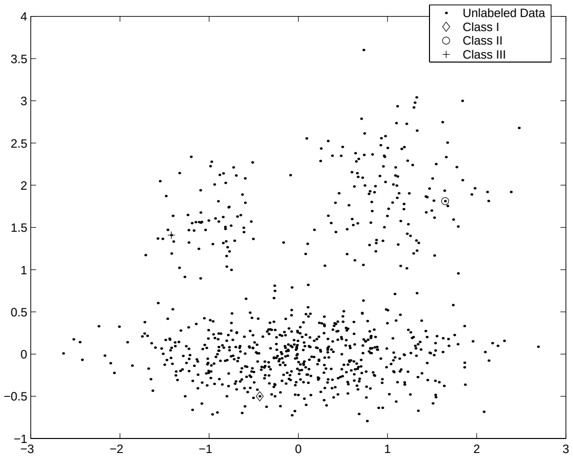

Figure 19. Training Data of 3 Gaussian Clusters

5. Conclusion

The guided cluster discovery problem is solved using attractive force-field formulations. The guided cluster discovery algorithm achieves integration of labeled information and spatial proximity information with force-field equations. Furthermore, the use of Markov assumptions allows a computationally feasible formulations to be developed. The guided cluster discovery approach excels in situations where a very small amount

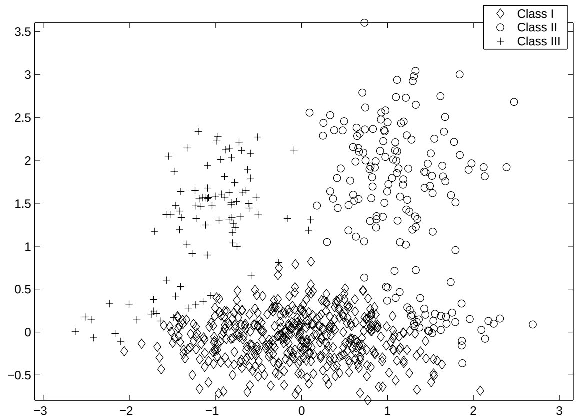

Figure 20. Nearest Neighbour classification of 3 Gaussian Clusters

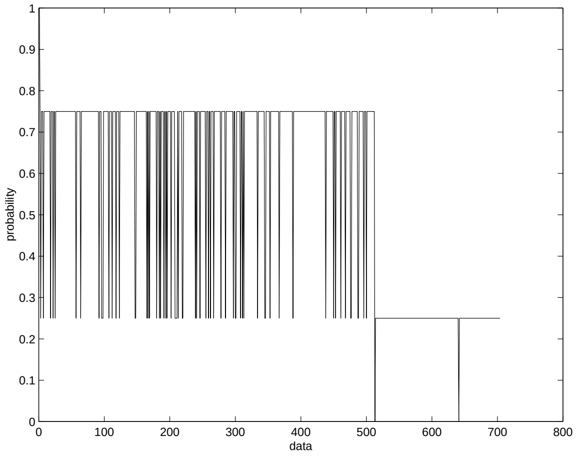

Figure 21. Initial probabilities for Class I, p^1

of labeled data and a large amount of unlabeled data is available for analysis. Experimental results shows that good clustering results can be obtained with only a few training samples as guidance and the spatial structure inherent in the problem can be readily utilized for improving the clustering process.

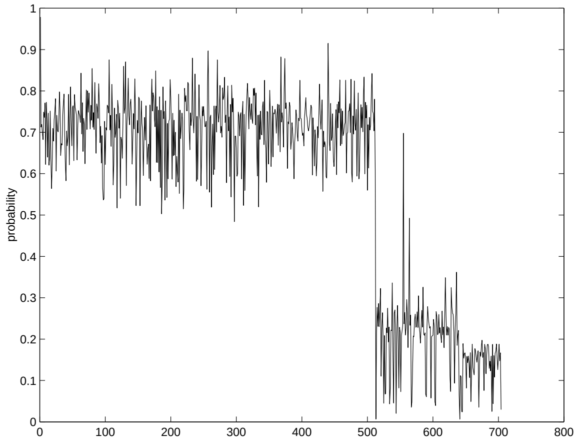

Figure 22. Final probabilities for Class I, p^1

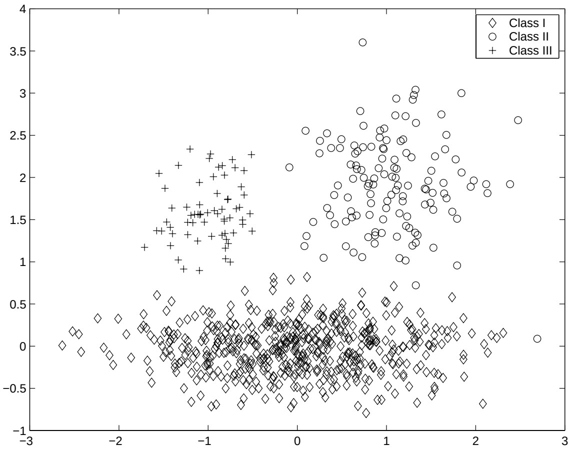

Figure 23. Final Estimated Labels

References

1. B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge,UK, 1996.
2. E. Forgy. Cluster analysis of multivariate data:efficiency vs. interpretablility of classifications. *Biometrics*, 21:768, 1965.
3. J. MacQueen. On convergence of k-means and partitions with minimum average variance. *Ann. Math. Statist.*, 36:1084, 1965.

4. J. Puzicha, T. Hofmann, and J. M. Buhmann. A theory of proximity based clustering: structure detection by optimization. *Pattern Recognition*, 33:617–634, 2000.

5. G. Karpis and E-H Han. Chameleon: Hierachical clustering using dynamic modeling. *IEEE Computer*, pages 68–75, August 1999.

6. R. Kannan, S. Vempala, and Veta A. On clusterings-good, bad and spectral. *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000.

7. A. Blum and Shuchi Chawla. Combining labeled and unlabeled data with co-training. In *The Eighteenth International Conference on Machine Learning*, 2001.

8. S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

9. K. Nigam, A. McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 34(1), 1999.

10. T. S. Chiang and Y. Chow. Optimization approaches to semi-supervised learning. In M. C. Ferris, O. L. Mangasarian, and J. S. Pang, editors, *Applications and Algorithms of Complementarity*. Kluwer Academic Publishers, 2000.

11. Ross Kindermann. *Markov random fields and their applications*. American Mathematical Society, Providence, R.I, 1980.

12. M. Nadler and E. P. Smith. *Pattern Recognition Engineering*. Wiley-interscience, New York, 1993.