Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

An Improved Sample Selection Algorithm in Fuzzy Decision Tree Induction

Ling-Cai Dong, Dan Wang, Xi-Zhao Wang
College of Mathematics and Computer Science,
Hebei University,
Baoding, China
wangxz@hbu.cn, dawingle@live.cn

***Abstract*—This paper improves a method of sample selection based on maximum entropy. Compared with the original method, the improved one takes the probability distribution of unlabeled instances into consideration. It selects the instances which can reduce the uncertainty of the whole unlabeled set to a great extent. The uncertainty reduction of the whole unlabeled set caused by an instance is measured by the instance's uncertainty and its influence index on the whole unlabeled set. To calculate the influence index conveniently, we introduces the similar matrix, the elements of which are the similarities measured by the distances between instances. The new method avoids the drawbacks that some abnormal or isolated samples may be selected by original method. Thus it can select the instances with more representation and more capability to resist noises. Our experimental results show that the performance of the classifier built from samples selected by the new algorithm is better than those selected by original method in the same time complexity.**

***Keywords*—Sample selection, probability distribution, similarity, classification ambiguity, fuzzy decision tree**

I. INTRODUCTION

Mass data are produced as the rapid and drastic technology development and the popularity of the Internet. There's no doubt that it is a hard task for human to process the large data effectively. Take an example in e-mail filtering. Thousands of e-mails are produced every day on the Internet. Undoubtedly, it is a tedious and boring work for people to classify so many e-mails. Take another example in the domain of life science. Abnormal structure of the proteins may cause the reduction and the loss of the biological activity, or even lead to disease, such as mad cow disease and Alzheimer. Therefore, the protein structure prediction and analysis is of great significance to the prevention and the treatment of the related disease. However, the predication of the proteinaceous structure is an extremely arduous and complex task. It will cost a great deal of time and energy. In reaction to the related phenomenon above, we hope that computer can deal with the task instead of human. In this way, experts only need to label a few instances for the classifier to learn, and a large proportion of unlabeled instances can be labeled by the classifier. How to select the instance as few as possible on the premise that the generalization capability of the classifier will not decrease, is becoming more and more important in machine learning.

The research on sample selection focuses three aspects[1]: uncertainty based methodology, version space based methodology and expectation error based methodology. Uncertainty-based methodology selects the instances with maximum uncertainty. Usually, these instances are situated in the neighborhood of the decision boundary and are hard to label, thus they are thought to be the most informative instances. This methodology can be applied to many induction learning, such as Logistic Regression[9], Hidden Markov Model (HMM)[10], Support Vector Machine (SVM) [11-12], uncertainty-based clustering[16], inductive logical programming [13] and decision tree[14] etc. Version space based methodology selects the instances that could reduce the version space at most. Usually, these instances could shrink the version space by half. The representative algorithms are QBC[2], SG net[3], QBag, QBoost [4], and Active Decorate[5] etc. Expectation error based methodology selects the instances that could reduce the expectation error at the utmost once they are selected and labeled. The methodology is applied to Native Bayes[6], Bayesian Network[7], Genetic Algorithm[17-18] and k-NN[8] etc. The methodology is thought to be best theoretically, for it directly takes the generalization error as the goal. However, too much calculation and time are needed.

This paper improves the maximum entropy based sample selection algorithm proposed in [14]. Comparing with original method, the improved algorithm takes the distribution of the unlabeled data into consideration. In stead of selecting the most uncertainty instance(s), it select the instance(s) which can probably cause the maximum average uncertainty reduction on the whole unlabeled set. The average uncertainty reduction on the whole unlabeled set caused by an instance is measured by the instance's uncertainty and its similarity to other unlabeled instances, in which the uncertainty and similarity are measured by classification ambiguity and Euclidean distance respectively. The experiments, conducted on UCI databases, show that the generalization capability of the improved algorithm is better than the original while no much time is cost.

The rest of the paper is organized as follows: section 2 gives some related basic concept and section 3 introduces the basic idea of improvement and the new algorithm. The experimental results conducted on UCI databases and the correspondence analyses are in section 4. Finally, we give the conclusion in section 5.

II. RELATED NOTIONS

Suppose that the universe of discourse $E = \{e_1, e_2, \cdots, e_N\}$ is made up by the training set T and testing set T', where

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

T and T' satisfy $T' \cap T = \varnothing, T' \cup T = X$. An instance e is described by the attribute set $A = \{A^{(1)}, A^{(2)}, \cdots, A^{(m)}\}$. Each attribute $A^{(k)} (1 \le k \le m)$ takes m_k values: $A_1^{(k)}, A_2^{(k)}, \cdots, A_{m_k}^{(k)}$ and there are R classes: $C_1, C_2, \cdots, C_R$. We designate an instance which has been labeled by expert as a sample denoted by a tuple (e_i, C_j) , which means the class label of the sample e_i is C_j .

Definition 1: Suppose that $\pi_1, \pi_2, \cdots, \pi_R$ is the classification possibility distribution of a sample e , then the classification ambiguity of the sample e is defined as

$$Ambig(e) = \sum_{i=1}^{R} (\pi_i^* - \pi_{i+1}^*) \ln i$$

where $\pi_{R+1}^* = 0$ and $\pi^* = \{\pi_1^*, \pi_2^*, \cdots, \pi_R^*\}$ is the normalized classification possibility distribution $(\pi_1, \pi_2, \cdots, \pi_R)$ with descending order.

Definition 2: A similar matrix $\mathbf{S} : |U| \times |U|$ is defined as

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1|U|} \\ \vdots & s_{ij} & \vdots \\ s_{|U|1} & \cdots & s_{|U||U|} \end{bmatrix}$$

where $|U|$ denotes the number of elements in unlabeled set and s_{ij} is the similarity degree between e_i and e_j . And the similarity degree can be calculated by

$$Similarity(e_i, e_j) = \exp(-D(e_i, e_j))$$

where $D(e_i, e_j) = \sqrt{\sum_{k=1}^{m} (A_{i_k}^{(k)} - A_{j_k}^{(k)})^2}$.

Definition 3: The classification ambiguity of an instance e_i will reduce to zero once it is labeled by expert. Then the uncertainty reduction of the instance after annotation is defined as

$$UR(e_i) = Ambig(e_i)_{bf} - Ambig(e_i)_{af}$$
$$= Ambig(e_i)_{bf}$$

where $Ambig(e_i)_{bf}$ is the ambiguity of the instance e_i before annotation and $Ambig(e_i)_{af}$ is the ambiguity after annotation, which equals to 0.

Without a doubt, the labeling process of the instance e_i will have the influences on the unlabeled instances especially those are in its neighborhood. Then we define the uncertainty reduction of e_j caused by the annotation of e_i as follows:

$$UR(e_j \mid e_i) = Similarity(e_j, e_i) \bullet UR(e_i)$$
$$= s_{ij} \bullet UR(e_i)$$

Then we can define the uncertainty reduction of the whole unlabeled set caused by the annotation of the instance e_i as

$$UR(U \mid e_i) = UR(e_i) \bullet \sum_{j=1}^{|U|} Similarity(e_j, e_i)$$
$$= UR(e_i) \bullet \sum_{j=1}^{|U|} s_{ij} = w_i \bullet UR(e_i)$$

in which w_i is called influence index of the instance e_i to the whole set and equals $\sum_{j=1}^{|U|} s_{ij}$.

$UR(U \mid e_i)$ can be viewed as the contribution of the instance e_i to the whole unlabeled set. It is the selection criterion of the new sample selection algorithm.

III. IMPROVED AMBIGUITY-BASED SAMPLE SELECTION ALGORITHM

A. Main idea

The uncertainty based sample selection methodology has been widely used in many domains as well as many machine learning algorithms [9-14]. Lewis & Gail propose the uncertainty sampling method with probability classifier in [9] which has been used to text classification. Based on the idea, [10] applies it to the learning of partially HMM (Hidden Markov Models), and [13] and [14] to natural language processing and fuzzy decision tree induction, respectively.

All the above methods select the most uncertainty samples for annotation. They consider the samples which close to the decision boundary with more information. Those samples which are far away from the decision boundary and can be easily labeled are thought with little or even no information for the current classifier.

However, there exist several drawbacks in the uncertainty based sample selection algorithms:

(1) The method ignores the instruction of the unlabeled data to current classifier. It only uses the labeled data to select samples, while the unlabeled set just plays the role of a pool for the classifier to select instances;

(2) The method may be sensitivity to noise for it select the instances closest to boundary which may be extremely particular and so without any representation;

(3) The method doesn't take the distribution of unlabeled set into consideration, so the isolated samples may be selected probably.

For the drawbacks mentioned above, we improve the maximum entropy based sample selection algorithm. Instead of selecting the instances with maximum uncertainty, the new algorithm selects the instances that could reduce the uncertainty

of the entire unlabeled set most. The main idea of the algorithm is described briefly below.

Firstly, construct the similar matrix of the unlabeled set. The matrix can be iteratively used without too much modification until the end of the algorithm. The elements of the matrix denote the similarity between instances. Secondly, train a classifier on the labeled set and predict the ambiguity of every unlabeled instance with the classifier. Thirdly, calculate the influence of every instance on the entire unlabeled set according to the similar matrix. Then get the classification ambiguity reduction of the whole unlabeled set caused by every unlabeled instance according to its ambiguity and influence index. Finally, select the one/ones which could reduce the classification ambiguity of the unlabeled set most for annotation and then add it/them to the labeled set with its true class and remove it/them at the same time from the unlabeled set. Repeat the procedure until the number of selected samples equals the predefined number.

From the description of the new algorithm's main idea, we can see that the criterion of the new algorithm to select samples concerns two aspects: the ambiguities of the instances and their influences to the unlabeled set once they are labeled. Compared with the original algorithm, the influence on the unlabeled set is considered. It introduces the similar matrix to describe the distribution of the unlabeled set and measure the influence on the labeled set of every unlabeled instance. Thus it can avoid the isolated instances to be selected and improve the robustness.

Here we take a simple example from [15] shown in Fig.1 to interpret the advantage of the new improved sample selection algorithm based on maximum uncertainty.

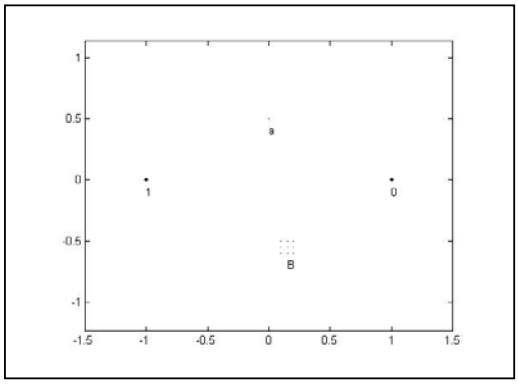

Figure 1. A simple example

Fig.1 shows a synthetic dataset with two labeled data (marked '1' and '0'), an unlabeled point 'a' lying in the center of the two labeled point in horizon, the coordinate of which is (0, 0.5) and a clustering 'B' including 9 unlabeled points, the coordinate of which are (0.1, -0.5), (0.15, -0.5), (0.2, -0.5), (0.1, -0.55), (0.15, -0.55), (0.2, -0.55), (0.1, -0.6), (0.15, -0.6), (0.2, -0.6) respectively from the top left one to the lower right one.

Obviously, the optimal decision surface obtained by the two labeled point is the vertical line $x = 0$ and the point 'a' is more uncertainty than every point in clustering B and has most uncertainty because it just lies on the decision boundary. So it will be considered as the most informative point by the sample selection algorithm based on maximum uncertainty and have the greatest contribution to current classifier on its prediction accuracy improvement. However, intuitively the points in clustering B should be more important than point 'a' because point 'a' is isolated from others with little representation and is likely to be a noise.

The improved new sample selection algorithm can solve the problem of avoiding such isolated data to be selected. From the figure we can see that the uncertainties are very similar among these unlabeled points, especially the left points of B and the point 'a'. But the points in clustering B are have more influence than point 'a' on the whole unlabeled set. Thus one or some points in clustering B can be selected instead of point 'a' by the criteria that is information amounts which is measured by the product of the uncertainty and the influence degree to the whole unlabeled set.

B. Algorithm description

The algorithm includes three components: an oracle G, a learner L and a data set X. Oracle G, which knows the label of all the data, may be a target function, a decision set or an expert in the correspondence domain etc. Learner L classifies all the unlabeled data and selects the most informative samples for current learner. Data set X includes labeled set X_L and unlabeled set X_U.

The goal of the algorithm is to get a learner L which can make $G(x) = L(x)$ for the elements $x \in X_U$ as many as possible.

At the beginning of the algorithm, there must be a learner/classifier, otherwise a labeled data set must be provided, which can be used to build a classifier. The steps of the algorithm are described below.

Step 1: Define a sample number to be selected.

Step 2: Build the similar matrix of unlabeled set.

Step 3: Train a fuzzy decision tree from the labeled set.

Step 4: Get the probability of every unlabeled instance belonging to every class by current decision tree. And then calculate the classification ambiguity of every unlabeled instance.

Step 5: Calculate the uncertainty reduction of the whole unlabeled set caused by of every unlabeled instance.

Step 6: Select the instance(s) which could reduce the uncertainty of the whole unlabeled set most for annotation. Then add it/them to the labeled set and remove it/them from the unlabeled set at the same time.

Step 7: Delete the correspondence row(s) and column(s) of the similar matrix.

Step 8: Judge whether the selected samples are enough. If not, go to step 3; otherwise go to step 9.

Step 9: Save the labeled set and train a fuzzy decision tree from the labeled set to predict unseen instances.

The algorithm flow chart is shown in Fig.2.

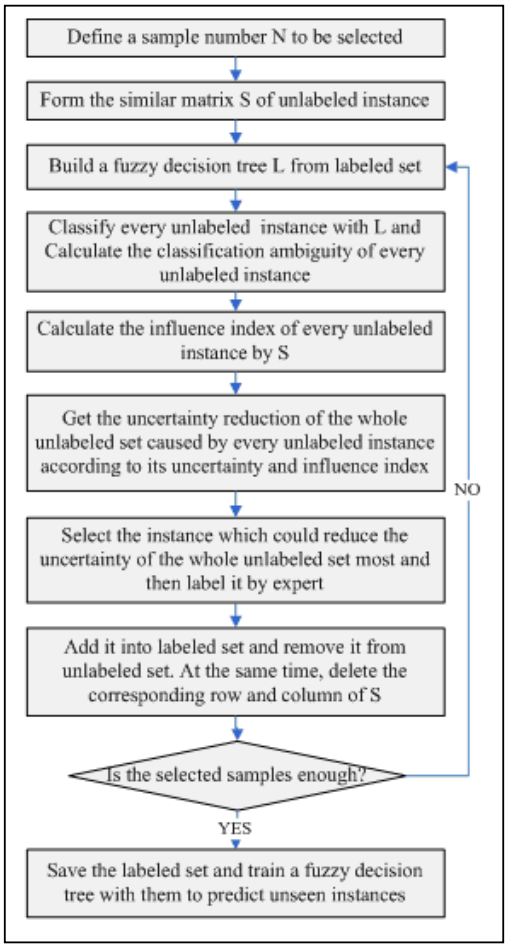

Figure 2. Flow chart of the improved sample selection algorithm

C. *Algorithm analysis*

The improved sample selection algorithm based on weighed maximum ambiguity introduces the similar matrix to calculate the influence index of every unlabeled instance on the whole unlabeled set and takes the most uncertain reduction of the unlabeled set as selection criteria, which is determined by two factors: influence index and classification ambiguity. The influence index of an instance to the whole unlabeled set is measured by the average similarities of the instance to every unlabeled instance. Thus the improved algorithm would not select abnormal instances with no representation.

The improved algorithm is not more complex and costs no more time than original because the similar matrix is built only once offline and has little change during the selection process. And most of the time consumption is the building the fuzzy decision tree. One drawback of the new algorithm is that it takes $O(n^2/2)$ storage more than original. But the storage will not greatly influence the execution of the program in these days.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We conduct the experiments on four databases of UCI. The information of the four databases is showed in Table 1.

TABLE I. DATABASES OF UCI

Databases	Attribute Number	Class Number	Instance Number
Iris	5	3	150
Wine	14	3	178
Glass	10	7	214
Sonar	61	2	208

In our experiments, all the numeric attributes is scattered into three clusters by k-means and fuzzified by triangular fuzzy number.

Firstly, we select 5% instances randomly from every database and all the others are thought to be unlabeled instances. Then we iteratively select one instance to label every time by three different ways: random selection, maximum entropy based selection and weighted maximum entropy based selection which is the improved method until the number of the selected samples equals to 20. The testing accuracies and elapsed time during the selection are recorded when a new instance is selected and added to the selected set.

The program is compiled with Matlab 7.1 and run on Pentium(R) 4 CPU 3.06GHz. The experiments are conducted 50 times on every database and the average testing accuracy and elapsed time are shown in Table 2.

TABLE II. EXPERIMENTAL RESULT

Databases		Iris	Wine	Glass	Sonar
Testing accuracy (%)	Random	0.9635	0.8120	0.5429	0.6870
	MABSS	0.9894	0.8349	0.5712	0.7194
	WMABSS	0.9898	0.8429	0.5755	0.7198
Selection time cost (seconds)	Random	0.0156	0.0156	0.0156	0.0156
	MABSS	0.0367	0.0453	0.0666	0.0606
	WMABSS	0.0415	0.0516	0.0634	0.0719

From Table 2, we can see that both of Maximum Ambiguity Based Sample Selection(MABSS) algorithm and the Weighted Maximum Ambiguity Based Sample Selection(MABSS) algorithm are better than random selection but cost more time. Comparing MABSS with WMABSS, the testing accuracy of decision tree trained from samples selected by WMABSS is higher than that of samples by MABSS although WMABSS costs more time than MABSS.

The advantage of WMABSS to MABSS and random selection during the selection process is also shown in Figs 3-6.

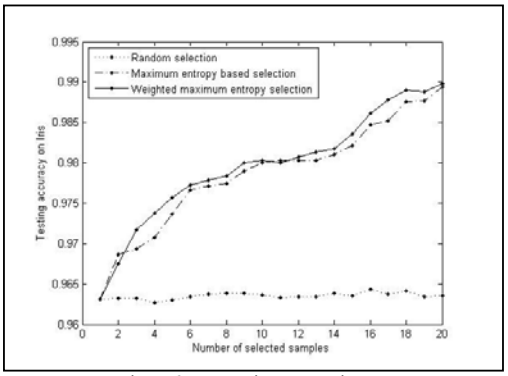

Figure 3. Experiment on Iris

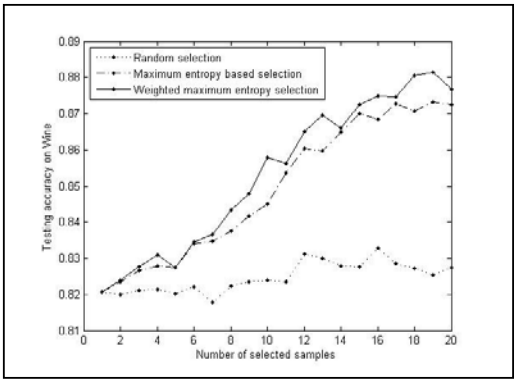

Figure 4. Experiment on Wine

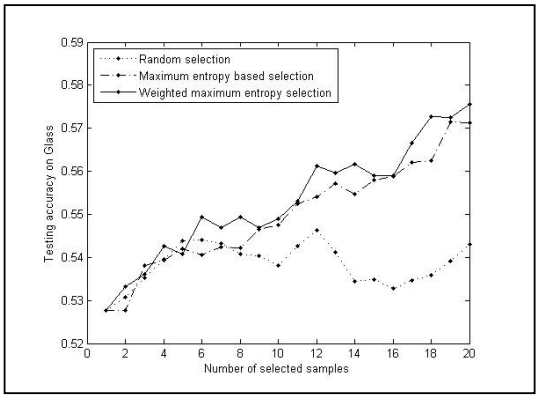

Figure 5. Experiment on Glass

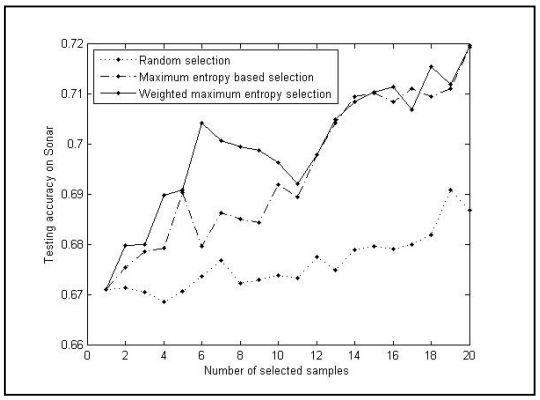

Figure 6. Experiment on Sonar

The figures clearly show that the improved sample selection method is better than both original method and random selection because the solid curves are always above the dashed curves and dotted curves.

V. CONCLUSION

This paper improves the maximum entropy based sample selection algorithm. It introduces the similar matrix to describe the probability distribution of the unlabeled instances and measure the influence of an unlabeled instance on the whole unlabeled set. The criterion to select samples is the uncertainty reduction of the whole unlabeled set caused by an unlabeled instance, which is determined by the classification ambiguity and the influence on the whole unlabeled set. The improved sample selection algorithm avoids the abnormal instances to be selected and is robust in noises while without time increasing. All of these have been shown in our experimental results conducted on UCI databases.

REFERENCES

[1] J. Long, J-P. Yin, E. Zhu, and W-T Zhao, "A survey of active learning," Journal of Computer Research and Development, vol. 45, pp. 300-304, 2008.

[2] H.S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," Proceedings of the fifth annual workshop on Computational Learning Theory, pp.287–294, 1992.

[3] D. Cohn and A.R. Ladner, "Improving generalization with active learning," Machine Learning, vol. 5, no. 2, pp. 201-221, 1994.

[4] N. Abe, H. Mamitsuka, "Query learning strategies using boosting and bagging," Proceedings of the 15th International Conference on Machine Learning, San Francisco, pp. 1-10, 1998.

[5] P. Melville and R.J. Mooney. "Diverse ensembles for active learning," Proceedings of the 15th International Conference on Machine Learning, San Francisco, 2004.

[6] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," Proceedings of the 15th International Conference on Machine Learning, San Francisco, pp. 441 – 448, 2001.

[7] S. Tong and D. Koller, "Active learning for parameter estimation in Bayesian networks," Proceedings of Advances in Neural Information, Cambridge, pp. 647-653, 2000.

[8] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers," Machine Learning, vol. 54, no. 2, pp. 125-152, 2004.

[9] D. Lewis and W.A. Gail. "A sequential algorithm for training text classifiers," Proceedings of the 17th ACM International Conference on

Research and Development in Information Retrieval, Berlin, pp. 3-12, 1994.

[10] T. Shcheffer and S. Wrobel. "Active learning of partially hidden Markov models," Proceedings of the ECML/PKDD, Berlin, 2001

[11] G. Schohn and D. Cohn, "Less is more: active learning with support vector machines," Proceedings of the 17th International Conference on Machine Learning; San Francisco, pp. 839-846, 2000.

[12] C. Campbell, N. Cristianini, and A. Smola, "A query learning with large margin classifiers," Proceedings of the 17th International Conference on Machine Learning; San Francisco, pp. 111 – 118, 2000.

[13] C. Tohompson, M.E. Califf, and R. Mooney, "Active learning for natural language parsing and information extraction," Proceedings of the 16th International Conference on Machine Learning; San Francisco, pp. 406 – 414, 1999.

[14] X. Wang, J.H. Yan, R. Wang, and C.R. Dong, "A sample selection algorithm in fuzzy decision tree induction and its theoretical analyses," IEEE International Conference on Systems, Man and Cybernetics, pp.3621-3626, October 2007.

[15] X. Zhu, "Semi-Supervised learning with graphs," Doctoral Thesis, May. 2005.

[16] Y. Hong and S. Kwong, "Learning assignment order of instances for constrained k-means clustering algorithm," IEEE Transactions on Systems Man and Cybernetics Part B, vol. 39, no. 2, pp.568-574, April 2009

[17] Y. Hong and S. Kwong, "Data clustering using virtual population based incremental learning algorithm with similarity matrix encoding strategy," ACM SIGEVO Genetic And Evolutionary Computation Conference (GECCO 2008), pp. 471-472, Nov. 2008

[18] Y. Hong and S. Kwong, "Genetic-guided semi-supervised clustering algorithm with instance-level constraints," ACM SIGEVO Genetic And Evolutionary Computation Conference (GECCO 2008), pp. 1381-1388, Nov. 2008