Pattern Recognition Letters 30 (2009) 298–305

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Resampling-based selective clustering ensembles

Yi Hong [a], Sam Kwong [a,*], Hanli Wang [a], Qingsheng Ren [b]

[a] *Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*
[b] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, PR China*

ARTICLE INFO

Article history:
Received 11 September 2007
Received in revised form 22 September 2008
Available online 22 October 2008

Communicated by M. Kamel

Keywords:
Clustering analysis
Clustering ensembles
Resampling technique

ABSTRACT

Traditional clustering ensembles methods combine all obtained clustering results at hand. However, we observe that it can often achieve a better clustering solution if only part of all available clustering results are combined. This paper proposes a novel clustering ensembles method, termed as resampling-based selective clustering ensembles method. The proposed selective clustering ensembles method works by evaluating the qualities of all obtained clustering results through resampling technique and selectively choosing part of promising clustering results to build the ensemble committee. The final solution is obtained through combining the clustering results of the ensemble committee. Experimental results on several real data sets demonstrate that resampling-based selective clustering ensembles method is often able to achieve a better solution when compared with traditional clustering ensembles methods.

1. Introduction

Clustering analysis classifies data items into groups such that items in the same group are more similar to each other, while they are more different in different groups. It has been known that clustering is an ill-posed combinatory optimization problem and no single algorithm is able to identify structures of all sorts of data sets (Jain et al., 1999; Duda et al., 2001). Numbers of clustering algorithms exist so far and some of them often produce contradictory clustering results, yet we can not claim which one is the best. In fact, almost all clustering algorithms are only valid for some data sets and may be invalid for other data sets. This uncertainty brings us a great difficulty to design a robust and stable clustering algorithm, therefore significantly limits applications of clustering analysis into more real-world data sets (Jain et al., 1999; Duda et al., 2001).

Clustering ensembles has been known as an effective method to improve the robustness and stability of clustering analysis (Strehl and Ghosh, 2002; Fred and Jain, 2005; Topchy et al., 2004a,b, 2005; Yang and Kamel, 2003). It works with combing multiple results of clustering algorithms of a data set without accessing its original features. Through leveraging the consensus across multiple clustering results, clustering ensembles gives a generic knowledge reuse framework for combining multiple clustering results (Strehl and Ghosh, 2002). Two crucial factors of clustering ensembles' success are as follows:

(1) To construct an accurate and diverse ensemble committee of the clustering ensembles.
(2) To design a proper consensus function to combine all clustering results of the ensemble committee.

Many recent studies have concentrated on the problem of constructing an accurate and diverse ensemble committee of the clustering ensembles'. Some of them used the same kinds of methods as the ones adopted in the area of supervised classifications, while others executed a clustering algorithm with different initializations, different parameters or different clustering criteria (Faceli et al., 2006). For example, Fred and Jain achieved a number of partitions with many diversities through running *K*-means clustering algorithm with random initializations and random numbers of clusters (Fred and Jain, 2005). Like the supervised classification area, bagging technique and resampling method were also adopted to generate a population of clustering results (Fischer and Buhmann, 2003; Monti et al., 2003). A random projection technique was introduced to build the ensembles of multiple clustering results for high dimensional data set by Fern and Brodley (2003, 2006). Resampling technique was used for constructing clustering ensembles in (Minaei-Bidgoli et al., 2004; Gondek and Hofmann, 2005; Topchy et al., 2004a). Several recent studies have demonstrated that all these methods are often able to construct accurate and diverse clustering ensembles (Strehl and Ghosh, 2002; Fred and Jain, 2005; Topchy et al., 2005; Fischer and Buhmann, 2003; Monti et al., 2003; Fern and Brodley, 2003, 2006; Strehl, 2002).

Up to the authors' best knowledge, most of the existing clustering ensembles methods combine all obtained clustering results at

* Corresponding author. Tel.: +852 2788 7704; fax: +852 2788 8614.
E-mail addresses: yihong@cityu.edu.hk (Y. Hong), CSSAMK@cityu.edu.hk (S. Kwong), ren-qs@cs.sjtu.edu.cn (Q. Ren).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.10.007

hand. However, we find that it can often achieve a better solution if only part of all available clustering results are used. We refer the kind of clustering ensembles method which uses only part of all obtained clustering results as the selective clustering ensembles. This paper deals with the selective clustering ensembles method. The inspirations of selective clustering ensembles method include two facets. Firstly, unlike classification problems where labels of data items are known beforehand, data items in unsupervised clustering problems are unlabeled. Therefore, we are unable to ensure that all clustering results are reliable. Thus, not all obtained clustering results can truly benefit for the final solution of clustering ensembles. From the standpoint of this view, it is reasonable to prune all these unusable clustering results before the ensembles of multiple clustering results. Secondly, clustering accuracies of all obtained clustering results may be significantly different. To roughly consider all of them contributing equally may weaken the quality of the final combined solution.

Selective ensembles method can also be traced back to the supervised classification area. In supervised classification area, it has been known that selective classifier ensembles can always achieve better solutions when compared with traditional ensemble methods (Ueda, 2000; Bandfield et al., 2005; Zhou et al., 2002; Zhang et al., 2006). A straightforward classifiers selection method is to rank the classifiers according to their individual performance on a held-out test set and pick the best ones (Caruana et al., 2004). Zhang et al. formulated the ensemble selection problem as a quadratic integer programming problem to look for a subset of classifiers that has the optimal accuracy-diversity trade-off (Zhang et al., 2006). In (Li et al., 2004), the authors proposed to use the clustering algorithm to prune redundant neural networks for maintaining the diversity of the ensemble committee of neural networks and very good results were reported. In unsupervised clustering area, however, data items are unlabeled beforehand. Therefore, it is unable to estimate the accuracy of a single clustering result through calculating its accuracy on the test set. To address the above problem, in this paper, we use the resampling technique to evaluate the quality of each clustering result. To the best knowledge of the authors', this is the first time to deal with the selective clustering ensembles method.

The remainder of this paper is arranged as follows. Section 2 demonstrates that clustering ensembles using part of all available clustering results is often able to achieve a better clustering solution when compared with traditional clustering ensembles methods that combine all of the obtained clustering results at hand. We describes resampling-based selective clustering ensembles method in Section 3. Our experimental results to demonstrate the potentials of resampling-based selective clustering ensembles are given in Section 4. Section 5 concludes this paper.

2. Clustering ensembles using part of all available clustering results

Before further illustrating about resampling-based selective clustering ensembles method, some notations used throughout this paper are given as follows. Let $D = \{d_1, d_2, \ldots, d_n\}$ denote a set of n data items without labels. A clustering result of the data set D can be represented as a label vector $I \in N^n$, where I_i is the label of the data item d_i. Let $C = \{I^{(1)}, I^{(2)}, \ldots, I^{(M)}\}$ be a set of M clustering results of the same data set D, where each $I^{(i)}$ is a label vector of $\{I_1^{(i)}, I_2^{(i)}, \ldots, I_n^{(i)}\}$. Clustering ensembles works to combine multiple clustering results $C = \{I^{(1)}, I^{(2)}, \ldots, I^{(M)}\}$ into a single consensus clustering result $I^{(\text{final})}$.

Traditional clustering ensembles methods utilize all of obtained clustering results at hand with the following steps employed: Step (1) A population of clustering results are obtained through executing different clusterers on the same data set. Step (2) The ensemble committee is constructed using all obtained clustering results. Step (3) A consensus function is adopted to combine all clustering results of the ensemble committee. Unlike classification problems where labels of data items are known beforehand, data items in unsupervised clustering problems are unlabeled. Therefore, there is no explicit correspondence between results provided by different clusterers. For example, the following two clustering results

$$I^{(1)} = (1,1,2,2,2,3,3) \quad I^{(2)} = (3,3,1,1,1,2,2)$$

are logically identical. To solve the problem of inconsistent clustering results, the label vector $I^{(i)}$ is firstly transformed into a similarity matrix $S^{(i)}_{n\times n}$ as follows:

$$S^{(i)}(j,k) = \begin{cases} 1 & \text{if } I_j^{(i)} = I_k^{(i)} \\ 0 & \text{if otherwise} \end{cases} \tag{1}$$

Accordingly, $\{S^{(1)}, S^{(2)}, \ldots, S^{(M)}\}$ can be obtained from M clustering results $\{I^{(1)}, I^{(2)}, \ldots, I^{(M)}\}$. After performing this transformation, all clustering results are consistent. Then clustering ensembles can be viewed as a process to combine all the similarity matrixes into a single consensus clustering result, that is

$$I^{(\text{final})} = \Omega(S^{(1)}, S^{(2)}, \ldots, S^{(M)}) \tag{2}$$

where Ω is a consensus function that is used to combine multiple clustering results into a single consensus one. A good ensemble committee should have high accuracy and many diversities (Fern and Brodley, 2003, 2006; Hadjitodorov et al., 2006; Kuncheva and Hadjitodorov, 2004). Considered that these two facets are sometimes contradictory, there must be a trade-off between the accuracy of a committee and the diversity of a committee. From the standpoint of this view, we get the following definition to compare any two ensemble committees:

Definition 1. Let $C^{(1)}$ and $C^{(2)}$ denote two ensemble committees. Their clustering accuracies are represented as $p^{(a)}_{c1}$ and $p^{(a)}_{c2}$. Their diversities are denoted as $p^{(d)}_{c1}$ and $p^{(d)}_{c2}$. If $p^{(a)}_{c1} > p^{(a)}_{c2}$ and $p^{(d)}_{c1} \geqslant p^{(d)}_{c2}$ or $p^{(a)}_{c1} \geqslant p^{(a)}_{c2}$ and $p^{(d)}_{c1} > p^{(d)}_{c2}$, we consider that the ensemble committee $C^{(1)}$ is better than the ensemble committee $C^{(2)}$.

Let $I^{(*)}$ denote the accurate partition of data set D and $S^{(*)}$ represent the similarity matrix of $I^{(*)}$, then average clustering accuracy of the ensemble committee $I^{(i)}$ can be calculated as follows:

$$p^{(a)} = \frac{\sum_{i=1}^{M} \|S^{(i)}, S^{(*)}\|}{M} \tag{3}$$

where $\|S^{(i)}, S^{(*)}\|$ is the agreement between two similarity matrixes $S^{(i)}$ and $S^{(*)}$. There have been many approaches to describing $\|\cdot\|$ such as rand index (Rand, 1971), adjusted rand index (Hubert and Arabie, 1985) and mutual information (Strehl and Ghosh, 2002). Apart from average clustering accuracy of an ensemble committee, the other important factor of a clustering ensemble's success is its diversity. We use average disagreement between any two clustering results in the ensemble committee to describe it as follows:

$$p^{(d)} = \frac{\sum_{j=1,\ldots,M} \sum_{k\neq j,k=1,2,\ldots,M} (1 - \|S^{(j)}, S^{(k)}\|)}{M(M-1)} \tag{4}$$

where $\|S^{(i)}, S^{(j)}\|$ is the agreement between two similarity matrixes $S^{(i)}$ and $S^{(j)}$.

According to Definition 1, the process of constructing a good ensemble committee can be viewed as a two-objective optimization problem as

$$\text{Objective 1}: \quad \max\left\{\frac{\sum_{i=1}^{M}\|S^{(i)},S^{(*)}\|}{M}\right\} \tag{5}$$

$$\text{Objective 2}: \quad \max\left\{1-\frac{\sum_{j\neq k,k,j,k=1,\ldots,M}\|S^{(j)},S^{(k)}\|}{M(M-1)}\right\} \tag{6}$$

We partition all obtained clustering results into two groups with respect to their accuracies as follows:

$$\{H^{a}=I^{(i)}|\|S^{(i)},S^{(*)}\|>p^{(a)},i=1,2,\ldots,M\} \tag{7}$$

and

$$L^{a}=\{I^{(i)}|\|S^{(i)},S^{(*)}\|\leqslant p^{(a)},i=1,2,\ldots,M\} \tag{8}$$

where $p^{(a)}$ is the average clustering accuracy of the ensemble committee. Likewise, according to the diversity of an ensemble committee, all obtained clustering results can be classified into two groups denoted as H^{d} and L^{d}:

$$H^{d}=\left\{I^{(i)}\middle|\sum_{j=1,\ldots,M,j\neq i}\|S^{(i)},S^{(j)}\|\leqslant\alpha,i=1,\ldots,M\right\} \tag{9}$$

and

$$L^{d}=\left\{I^{(i)}\middle|\sum_{j=1,\ldots,M,j\neq i}\|S^{(i)},S^{(j)}\|>\alpha,i=1,\ldots,M\right\} \tag{10}$$

where

$$\alpha=(M-1)\cdot(1-p^{(d)})$$

$p^{(d)}$ is defined by the formula (4). We define the set R as follows:

$$R=L^{a}\cap L^{d}. \tag{11}$$

Definition 2. Let $\widetilde{I}$ be an element of the whole clustering results C, $C\setminus\{\widetilde{I}\}$ represents a subset of C through removing the element $\widetilde{I}$ from C.

Then the following theorem can be obtained:

Theorem 1. *If R is not an empty set, there must be a subset C' of C such that to combine the clustering results of C' is better than to combine all available clustering results of C.*

Proof. If R is not an empty set, to randomly pick one element $\widetilde{I}$ from R. Considered that

$$R\subset C$$

then

$$\widetilde{I}\in C$$

The average clustering accuracy $\tilde{p}^{(a)}$ of $C\setminus\{\widetilde{I}\}$ can be calculated as

$$\tilde{p}^{(a)}=\frac{\sum_{i=1}^{M}\|S^{(i)},S^{(*)}\|-\|\tilde{S},S^{(*)}\|}{M-1}$$

where $\tilde{S}$ is the similarity matrix of $\widetilde{I}$. According to the definition of R

$$\|\tilde{S},S^{(*)}\|<p^{(a)}$$

and

$$\sum_{i=1}^{M}\|S^{(i)},S^{(*)}\|=M\cdot p^{(a)}$$

We obtain

$$\tilde{p}^{(a)}>p^{(a)} \tag{12}$$

The diversity $p^{\widetilde{(d)}}$ of the ensemble committee $C\setminus\{\widetilde{I}\}$ can be calculated as follows:

$$\tilde{p}^{(d)}=1-\frac{\left(\sum_{j=1,\ldots,M}\sum_{k\neq j,k=1,2,\ldots,M}\|S^{(j)},S^{(k)}\|-\beta\right)}{(M-1)(M-2)}$$

where

$$\beta=2\cdot\sum_{i=1,\tilde{S}\neq S^{(i)}}^{M}\|\tilde{S},S^{(i)}\|$$

According to the definition of R

$$\beta\geqslant 2\cdot(M-1)\cdot(1-p^{(d)})$$

and

$$\sum_{j=1,\ldots,M}\ \sum_{k\neq j,k=1,2,\ldots,M}\|S^{(j)},S^{(k)}\|=M\cdot(M-1)\cdot(1-p^{(d)})$$

Then we can obtain the following inequality

$$\tilde{p}^{(d)}\geqslant p^{(d)} \tag{13}$$

From inequalities 12, 13 and Definition 1, we can get the conclusion that to combine the clustering results of $C\setminus\{\widetilde{I}\}$ is better than to combine the whole clustering results C.

Theorem 1 tells us that it may achieve a better solution if only part of all available clustering results are combined. Here we give a simple example to illustrate this point. Given the clustering results $C=\{I^{(1)},I^{(2)},I^{(3)},I^{(4)},I^{(5)},I^{(6)}\}$ as

$$I^{(1)}=(2,2,2,1,1,1),\quad I^{(2)}=(2,2,1,2,1,1)$$
$$I^{(3)}=(1,1,2,2,2,2),\quad I^{(4)}=(1,1,2,1,2,2)$$
$$I^{(5)}=(1,1,1,1,2,2),\quad I^{(6)}=(1,1,2,1,2,2)$$

and one of accurate clustering solutions[1] is $I^{(*)}=(1,1,1,2,2,2)$. Accuracies of all obtained clusterers are 100% for $I^{(1)}$, 46.7% for $I^{(2)}$, 66.7% for $I^{(3)}$, 46.7% for $I^{(4)}$, 66.7% for $I^{(5)}$, 46.7% for $I^{(6)}$ and their average clustering accuracy equals to 62.5%. Values of $\frac{\sum_{j=1,j\neq i}^{6}\|S^{(i)},S^{(j)}\|}{5}$ are 54.7% for $I^{(1)}$, 77% for $I^{(2)}$, 67% for $I^{(3)}$, 77% for $I^{(4)}$, 67% for $I^{(5)}$ and 77% for $I^{(6)}$. According to formula (7)–(11), we can obtain $R=\{I^{(2)},I^{(4)},I^{(6)}\}$. Therefore, to combine the clustering results of $C\setminus\{I^{(2)}\}$ or $C\setminus\{I^{(4)}\}$ or $C\setminus\{I^{(6)}\}$ is better than to combine the whole clustering results of C.

It is noted the condition that R is not empty is not necessary for selective clustering ensembles to achieve better solutions than the classical clustering ensemble methods. If R is empty, whether selective clustering ensembles is better than classical clustering ensembles or not is dependent on the trade-off between the accuracy of the committee and the diversity of the committee. However, our experimental results on several real data sets have demonstrated that selective clustering ensembles is often able to achieve a better solution when compared with traditional clustering ensembles methods.

3. Resampling-based selective clustering ensembles

The above section has demonstrated that clustering ensembles using part of all obtained clustering results can often be better than those using the whole clustering results. In this section, we describe the framework of resampling-based selective clustering ensembles that selectively combines part of all obtained clustering results. Resampling-based selective clustering ensembles works with evaluating the qualities of all obtained clustering results

[1] To classify data items into k groups, the overall number of accurate clustering solutions is $k!$.

using resampling technique and choosing part of promising clustering results to build the ensemble committee. The final solution is achieved through combining all the selected clustering results of the ensemble committee.

Algorithm 1. Framework of resampling-based selective clustering ensembles.

```
(1) $\{I^{(1)}, I^{(2)}, \dots, I^{(M)}\} \leftarrow$ A population of clustering results are obtained;
(2) $\{q^{(a)}_{(1)}, q^{(a)}_{(2)}, \dots, q^{(a)}_{(M)}\} \leftarrow$ Calculate accuracies of all results by resampling
(3) $\{q^{(d)}_{(1)}, q^{(d)}_{(2)}, \dots, q^{(d)}_{(M)}\} \leftarrow$ Calculate individual diversity factors as $q^{(d)}_{(i)} = 1 - \frac{\sum_{j=1, j\neq i}^{M} \|S^{(i)}, S^{(j)}\|}{M-1} i = 1, \dots, M$;
(4) $\{\mathrm{fitness}(I^{(1)}), \dots, \mathrm{fitness}(I^{(M)})\} \leftarrow$ Calculate fitness values as: $\mathrm{fitness}(I^{(i)}) = (1-\lambda) \cdot \frac{q^{(a)}_{(i)}}{Q^a} + \lambda \cdot \frac{q^{(d)}_{(i)}}{Q^d} i = 1, \dots, M$;
(5) $\{'I^{(1)}, 'I^{(2)}, \dots, 'I^{(M)}\} \leftarrow$ Reorder clustering results such that: $\mathrm{fitness}('I^{(1)}) \geqslant \cdots . \geqslant \mathrm{fitness}('I^{(M)})$;
(6) $\{'I^{(1)}, 'I^{(2)}, \dots, 'I^{(N)}\}(N < M) \leftarrow$ Choose $N$ best clustering results;
(7) $I^{(\mathrm{final})} \leftarrow$ Combine $\{'I^{(1)}, 'I^{(2)}, \dots, 'I^{(N)}\}$ together.
```

As illustrated in the above section, it is a two-objective optimization problem to construct a good ensemble committee. We expect that the obtained ensemble committee has a high clustering accuracy and many diversities. The diversity of an ensemble committee can be described as follows:

$$p^{(d)} = \frac{\sum_{j=1,\dots,M} \sum_{k \neq j, k=1,2,\dots,M} (1 - \|S^{(j)}, S^{(k)}\|)}{M(M-1)}$$

where $\|\cdot\|$ is the agreement between two similarity matrixes. Another version of this formula is as follows:

$$p^{(d)}_{(i)} = \frac{\sum_{i=1}^{M} q^{(d)}_{(i)}}{M} \tag{14}$$

where $q^{(d)}_{(i)}$ can be calculated as

$$q^{(d)}_{(i)} = 1 - \frac{\sum_{j=1, j\neq i}^{M} \|S^{(i)}, S^{(j)}\|}{M-1} \quad i = 1, \dots, M \tag{15}$$

We call $q^{(d)}_{(i)}$ as the individual diversity factor of the clustering result $I^{(i)}$. The clustering result $I^{(i)}$ with a high $q^{(d)}_{(i)}$ means that the clustering result $I^{(i)}$ has a large difference from other obtained clustering results.

For unsupervised clustering, all data items are unlabeled beforehand. Therefore, we are unable to calculate the accuracy of a clustering result by its class labels of items. One feasible approach to estimate the quality of a clustering result is resampling technique (Levine and Domany, 2001; Tseng and Wong, 2003). Resampling technique evaluates the quality of a clustering result $I^{(i)}$ with the following steps employed: (a) T subsets of the full data set $\{R^1, R^2, \dots, R^T\}$ are randomly selected from the full data set; (b) T clustering results $\{J^{(1)}, J^{(2)}, \dots, J^{(T)}\}$ are obtained through executing the i^{th} clusterers[2] on these T different subsets of the full data set; (c) The quality $q^{(a)}_{(i)}$ of the clustering result $I^{(i)}$ can be estimated as

$$q^{(a)}_{(i)} = \frac{\sum_{k=1}^{T} \|I^{(i)}, J^{(k)}\|_{R^k}}{T} \tag{16}$$

where $\|I^{(i)}, J^{(k)}\|_{R^k}$ is the agreement between the clustering result $I^{(i)}$ and the clustering result $J^{(k)}$ on the selected subset R^k of the full data set.

After $q^{(a)}_{(i)}$ and $q^{(d)}_{(i)}$ have been successfully calculated, we use the method proposed by Optiz to balance between the accuracy and the diversity (Opitz, 1999). The accuracy $q^{(a)}_{(i)}$ and the diversity $q^{(d)}_{(i)}$ are combined together as the fitness value of the clustering result $I^{(i)}$

$$\mathrm{fitness}(I^{(i)}) = (1-\lambda) \cdot \frac{q^{(a)}_{(i)}}{Q^a} + \lambda \cdot \frac{q^{(d)}_{(i)}}{Q^d} \tag{17}$$

where $\lambda (0 \leqslant \lambda \leqslant 1)$ is a control parameter that is used to balance between accuracy and diversity

$$Q^a = \max\{q^{(a)}_1, q^{(a)}_2, \dots, q^{(a)}_M\} \tag{18}$$

and

$$Q^d = \max\{q^{(d)}_1, q^{(d)}_2, \dots, q^{(d)}_M\} \tag{19}$$

After fitness values of all obtained clustering results have been calculated, we rank all the clustering results according to their fitness values and select the $N (N < M)$ best ones to build the ensemble committee. The final clustering solution is obtained through combining all the clustering results of the ensemble committee. Algorithm 1 gives the framework of resampling-based selective clustering ensembles method.

4. Experimental results

Eight real data sets are selected to test the performance of resampling-based selective clustering ensembles method from UCI Machine Learning Repository (Blake and Merz, 1998). Their names and characteristics are given in Table 1. In our experiments, we use the Rand Index method to describe the agreement between any two clustering results $I^{(i)}$ and $I^{(j)}$, that is

$$\|I^{(i)}, I^{(j)}\| = \frac{2 \cdot (n_{00} + n_{11})}{n \cdot (n-1)}$$

where n_{11} is the number of pairs of items that are both in the same group in $I^{(i)}$ and also both in the same group in $I^{(j)}$ and n_{00} denotes the number of pairs of items that are in different groups in $I^{(i)}$ and also in different groups in $I^{(j)}$. The same measure is also adopted to measure the accuracy of the final clustering solution obtained by clustering ensembles methods. k-means clustering algorithm is adopted as the "base" clustering algorithm because of its wide application in constructing clustering ensembles. A population of clustering results is obtained through randomly picking half of features from the data set with replacement and partitioning data items according to these selected features. The above steps are independently executed for 100 runs and 100 clustering results are obtained. A popular clustering ensembles approach proposed by Strehl and Ghosh[3] is adopted to combine these clustering results (Strehl, 2002). In (Strehl, 2002), the authors in fact proposed three ensemble approaches: CSPA, HGPA and MCLA to combine multiple clustering results of the same data set. Three final clustering results can therefore be achieved by CSPA, HGPA and MCLA, respectively and the one with the maximum average normalized mutual information is returned as the final clustering result. As the parameters used in (Levine and Domany, 2001), the value T of the number of resampled subsets is set as 50 and each subset of the full data set includes 60% of the full data set. All experiments are executed for 20 independent runs and their results are averaged and reported.

Firstly, we fix the number of selected clustering results to 50, the control parameter λ to 0.5 and compare clustering accuracies of resampling-based selective clustering ensembles method and traditional clustering ensembles methods. Their results are given

[2] The clustering result $I^{(i)}$ is obtained by the ith clusterers.

[3] We would like to thank the authors to put the code online: http://www.lans.ece.utexas.edu/strehl/soft.html.

Table 1
Data set and their characteristics.

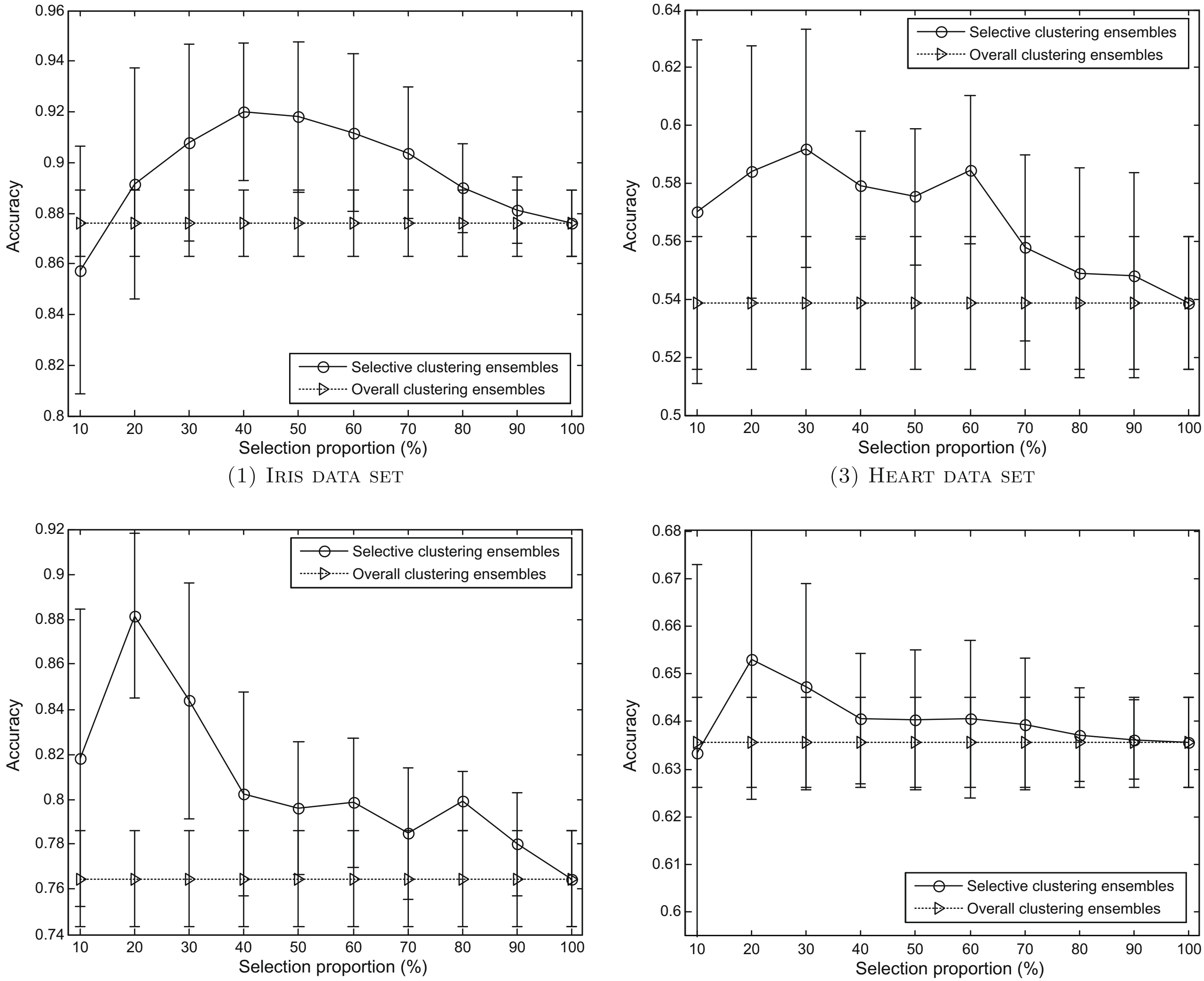
RANK	DATA SET	DATA ITEM	FEATURE	CLASS
1	IRIS	150	4	3
2	WINE	178	13	3
3	HEART	270	13	2
4	LUNG	32	56	3
5	WDBC	569	30	2
6	VEHICLE	846	18	4
7	SEGMENTATION	2310	19	7
8	SAT. IMAGE	6435	36	6

Table 2
Clustering accuracies obtained by traditional clustering ensembles methods and selective clustering ensembles method.

DATA SET	TA. CLU. ENS. (%)	SEL. CLU. ENS. (%)
IRIS	87.6 ± 1.3	91.8 ± 2.7
WINE	76.3 ± 2.1	79.1 ± 2.9
HEART	53.8 ± 2.8	57.6 ± 2.4
LUNG	63.4 ± 0.6	64.5 ± 1.6
WDBC	75.4 ± 0.3	78.9 ± 0.9
VEHICLE	63.6 ± 0.2	64.3 ± 0.3
SEGMENTATION	87.2 ± 0.6	89.0 ± 0.6
SAT. IMAGE	83.4 ± 0.4	85.9 ± 0.5

in Table 2. From Table 2, we can observe that resampling-based selective clustering ensembles method achieves comparative or better solutions for all tested data sets when compared with traditional clustering ensembles methods. For example, resampling-based selective clustering ensembles method achieves around 91.8% accuracy for Iris data set and 57.6% accuracy for Heart data set. Both of them are significantly higher than those 87.6% for Iris data set and 53.8% for Heart data set obtained by traditional clustering ensembles methods. The superiority of selective clustering ensembles method in clustering accuracy lets us know that not all available clustering results are useful for constructing the ensemble committee and pruning these unusable clustering solutions before combining them into the final clustering result can often improve the quality of the final clustering solution of clustering ensembles.

Secondly, we test the effect of selection proportion through increasing selection proportion from 0.1 to 1.0 and calculating accuracies of selective clustering ensembles method. Fig. 1 gives the results. The following two phenomena can be observed from Fig. 1. First of all, clustering ensembles using part of all available clustering results can achieve better solutions when compared with clustering ensembles using all available clustering results. The other phenomenon observed from Fig. 1 is the performance of selective clustering ensembles method is closely related with



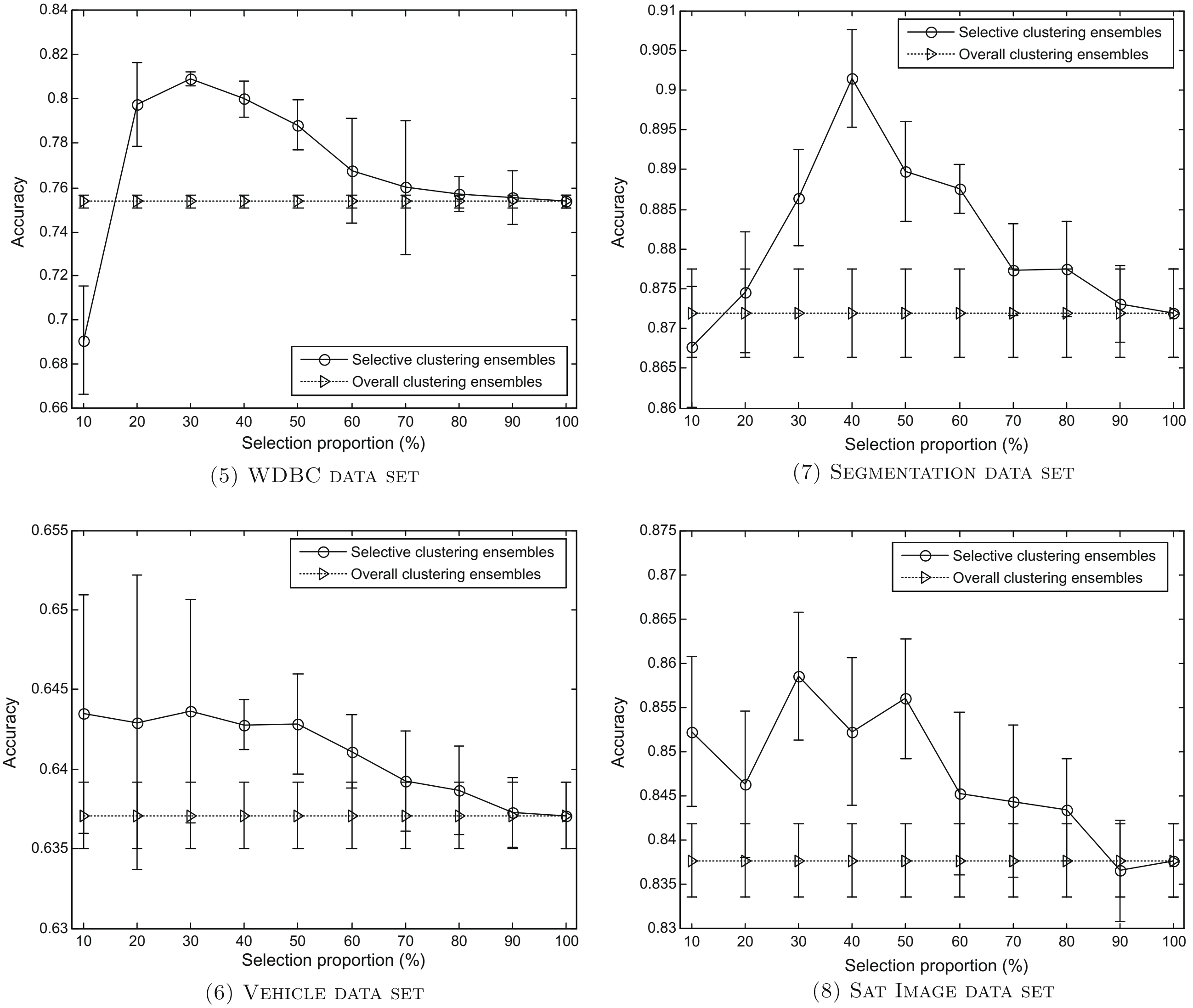
(1) Iris data set

(3) Heart data set

(2) Wine data set

(4) Lung data set

Fig. 1. Clustering accuracy under different selection proportions. Graph based clustering ensembles method was proposed by Strehl and Ghosh in (Strehl, 2002).



(5) WDBC data set

(7) Segmentation data set

(6) Vehicle data set

(8) Sat Image data set

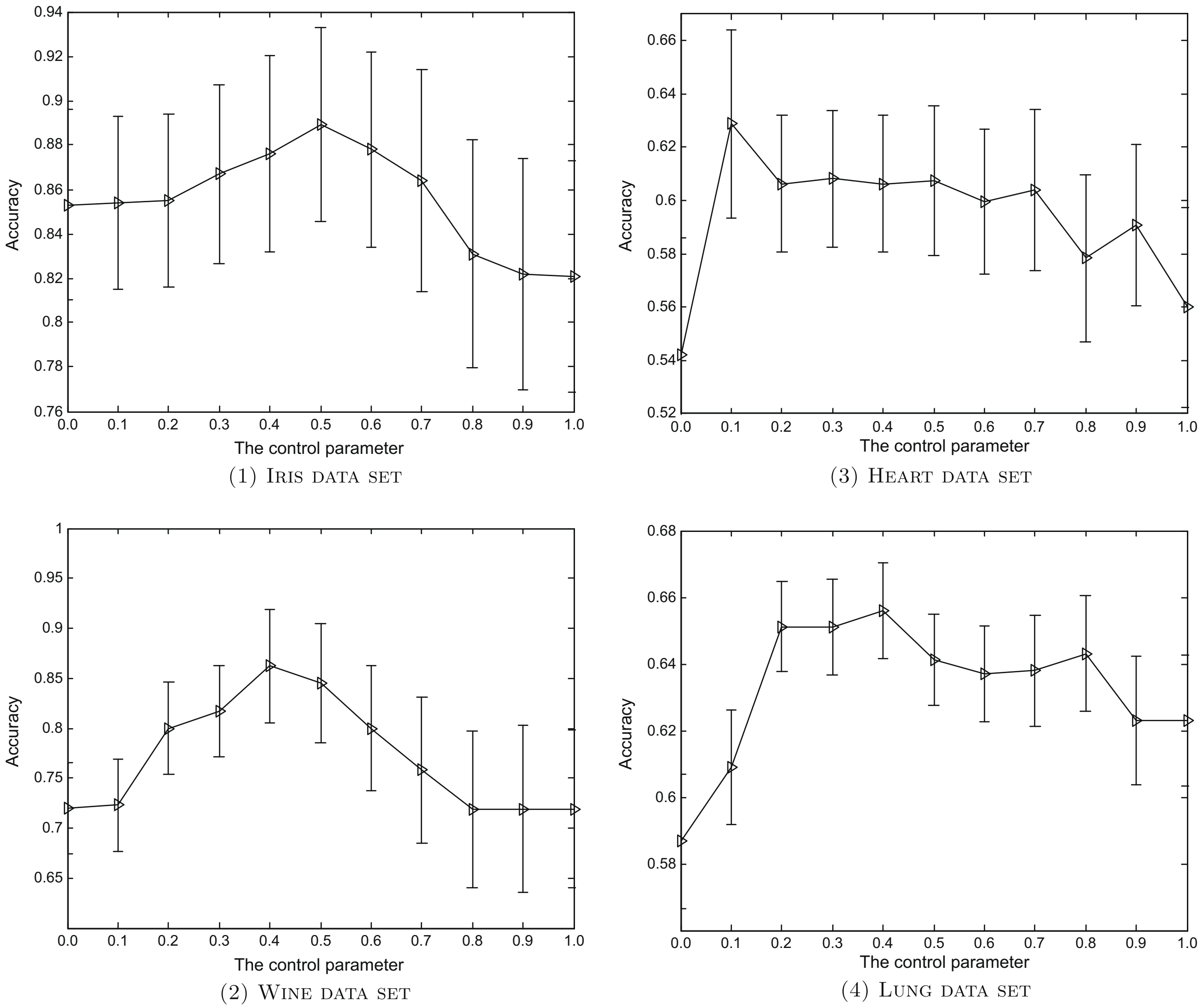
Fig. 1 (*continued*)

the value of selection proportion. For example, selective clustering ensembles method can achieve 85.7% accuracy for Iris data set if its selection proportion equals to 0.1. Its accuracy increases up to 91.8% if its selection proportion is set as 0.5. However, the accuracy of selective clustering ensembles method decreases to 87.6%, if the selection proportion increases up to 1.0. Moreover, a good selection proportion of selective clustering ensembles method varies from a data set to a data set. For example, a good selection proportion is 0.4 for Iris data set, 0.3 for WDBC data set, 0.1 for Vehicle data set and 0.3 for Sat Image data set based on the experimental results in Fig. 1.

Finally, we fix the selection proportion to 0.3 and test the effect of the control parameter λ through increasing its value from 0.0 to 1.0. It should be noted that the control parameter λ is used to balance between accuracy and diversity for constructing a good ensemble committee of clustering ensembles. The value 0.0 of the control parameter λ means that only clustering accuracy is employed for selectively constructing clustering ensembles, while the value 1.0 of the control parameter λ represents that only diversity is considered. The results are given in Fig. 2. From Fig. 2, we can see that the performance of selective clustering ensembles method is closely related with the control parameter λ and a moderate value of the control parameter can often lead to good clustering ensembles. For example, selective clustering ensembles with the control parameter 0.3 can achieve 60% accuracy for Heart data set, that is significantly higher than 54.5% accuracy when the control parameter equals 0.0 and 56.2% accuracy when the control parameter equals 1.0. These results tell us that a good ensemble committee should draw a good balance between accuracy and diversity and different data sets have different good control parameters.

5. Conclusions

In this paper, a novel selective clustering ensembles method is proposed. We have proved in theory that the proposed selective clustering ensembles method can be better than traditional clustering ensembles methods in solving unsupervised classification problems. To achieve this, the resampling technique is employed for the first time by the proposed approach to select part of all obtained clustering results to construct the ensemble committee. Experimental results on several real data sets have demonstrated that the proposed selective clustering ensembles method is able to achieve a better clustering performance than traditional clustering ensembles methods.



(1) Iris data set

(3) Heart data set

(2) Wine data set

(4) Lung data set

Fig. 2. Clustering accuracy under different control parameters.

Acknowledgements

The work was partially supported by City University Strategic Grant No. 7002294 and a grant from the Research Grants Council of Hong Kong Special Administrative Region, China Project No. 9041236/CityU 114707. The authors would like to thank the comments and suggestions from the reviewers.

References

Bandfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P., 2005. Ensemble diversity measures and their application to thinning. J. Informat. Fusion 5, 49–62.

Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases. Available from: <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2004. Ensemble selection from libraries of models. In: Proc Internat. Conf. on Machine Learning, pp. 18–25.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classif.

Faceli, K., de Carvalho, A., de Souto, N., 2006. Multi-objective clustering ensemble. In: Internat. Conf. Hybrid Intelligent Systems, pp. 51–51.

Fern, X.Z., Brodley, C.E., 2003. Random projection for high dimensional data clustering. In: Proc. Internat. Conf. on Machine Learning, pp. 186–193.

Fern, X.Z., Brodley, C.E. 2006. Clustering ensembles for high dimensional clustering: An empirical study. Technical report CS06-30-02, Oregon State University.

Fischer, B., Buhmann, J., 2003. Bagging for path-based clustering. IEEE Trans. Pattern Anal. Machine Intell. 25, 1411–1415.

Fred, A., Jain, A.K., 2005. Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Machine Intell. 27, 835–850.

Gondek, D., Hofmann, T., 2005. Non-redundant clustering with conditional ensembles. In: ACM SIGKDD Internat. Conf. on Knowledge Discovery in Data Mining, pp.70–77.

Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P., 2006. Moderate diversity for better cluster ensembles. J. Informat. Fusion 7, 264–275.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Comput. Surv. 13, 264–323.

Kuncheva, L.I., Hadjitodorov, S.T., 2004. Using diversity in cluster ensemble. In: Proc. IEEE Internat. Conf. on System, Man and Cybernetics, pp. 1214–1219.

Levine, E., Domany, E., 2001. Resampling method for unsupervised estimation of cluster validity. Neural Comput. 13, 2573–2593.

Li, G., Yang, J., Kong, A.S., Chen, N., 2004. Clustering algorithm based selective ensemble. J. Fudan Univ. 43, 689–695.

Minaei-Bidgoli, B., Topchy, A., Punch, W.F., 2004. Ensembles of partitions via data resampling. In: Internat. Conf. on Information Technology: Coding and Computing, p. 188.

Monti, S., Tamaya, P., Mesirov, J., Golub, T., 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression moscovery data. J. Mach. Learn. 52, 91–118.

Opitz, D., 1999. Feature selection for ensembles. In: Proc. National Conf. on Artifical Intelligence, vol. 7, pp. 379–384.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66, 846–850.

Strehl, A., 2002. Relationship-based clustering and cluster ensembles for high-dimensional data mining. Dissertation of The University of Texas at Austin.

Strehl, A., Ghosh, J., 2002. Clustering ensemble – a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617.

Topchy, A., Minaei, B., Jain, A.K., Punch, W., 2004a. Adaptive clustering ensembles. In: Internat. Conf. on Pattern Recognition, pp. 272–275.

Topchy, A., Law, M., Jain, A.K., Fred, A., 2004b. Analysis of consensus partition in cluster ensemble. In: Internat. Conf. on Data Mining, pp. 225–232.

Topchy, A., Jain, A.K., Punch, W., 2005. Clustering ensembles: Models of consensus and weak partitions. IEEE Trans. Pattern Anal. Machine Intell. 27, 1866–1881.

Tseng, G.C., Wong, W.H., 2003. A method for tight clustering: With application to microarray. In: Proc. IEEE Computer Society Bioinformatics Conf., pp. 396–397.

Ueda, N., 2000. Optimal linear combination of neural networks for improving classification performance. IEEE Trans. Pattern Anal. Machine Intell., 207–215.

Yang, Y., Kamel, M., 2003. Clustering ensemble using swarm intelligence. In: IEEE Swarm Intelligence Sysposium, pp. 65–71.

Zhang, Y., Burer, S., Street, W., 2006. Ensemble pruning via semi-definite programming. J. Mach. Learn. Res. 7, 1315–1338.

Zhou, Z., Wu, J., Tang, J., 2002. Ensembling neural networks: Many could be better than all. Artif. Intell. 137, 239–263.