Selective Clustering Ensemble Based on Covariance

Xuyao Lu, Yan Yang*, and Hongjun Wang

School of Information Science & Technology
Southwest Jiaotong University
Chengdu, 610031, P.R. China
lxy@my.swjtu.edu.cn, {yyang,wanghongjun}@home.swjtu.edu.cn

Abstract. Clustering Ensemble effectively improves clustering accuracy, stability and robustness, which is most resulted from the diversity of the base clustering results. It is a key point to measure the diversity of clustering results. This paper proposes a method to measure diversity of base clustering results and a covariance-based selective clustering ensemble algorithm. Experiments on 20 UCI data sets show that this algorithm effectively improves the clustering performance.

Keywords: Clustering ensemble, Covariance, Selective ensemble.

1 Introduction

Clustering is the process of splitting the set of physical or abstract objects into similar object classes [1]. It is a high similarity between the same class and a great difference between the different classes. Ensemble learning gets base learners by different methods, and obtains a final results by combining base learners in some way [2]. Strehl et al. [3] proposed the clustering ensemble in 2002. Clustering ensemble is a method of aggregating the multiple division collection of one object into a final clustering result. Clustering ensemble effectively reduces the impact on noise and outliers, and increases the clustering stability and robustness.

Recent years, there are many research works in clustering ensemble. Topchy et al. [4] designed a mixture model for clustering ensemble, and they offered a probabilistic model of consensus with a finite mixture of multinomial distributions in a space of clustering. A new consensus function by the generalized mutual information was proposed in [5]. Luo et al. [6] used information theory to design a genetic algorithm to combine multiple clusterings. Hassan et al. [7] developed a ensemble method with majority voting and parallel fusion in conjunction with a neural classifier. Mohammadi et al. [8] stated an evolutionary approach to clustering ensemble, and they used an evolutionary combinational clustering method to find the number of clusters. Iqbal et al. [9] proposed the semi-supervised clustering ensemble by voting, and they introduced a flexible two parameters weighting mechanism in their algorithm. The semi-supervised

* Corresponding author.

Z.-H. Zhou, F. Roli, and J. Kittler (Eds.): MCS 2013, LNCS 7872, pp. 179–189, 2013.

cluster ensemble model based on bayesian network was designed. And the variational inference oriented semi-supervised cluster ensemble is illustrated in this paper [10]. Jia et al. [11] presented a bagging-based spectral clustering ensemble selection. Yang et al. [12] presented a semi-supervised clustering ensemble based on multi-ant colonies, and they incorporated pairwise constraints not only in each ant colony clustering process, but also in computing new similarity matrix. Iam-On et al. [13] advanced a link-based cluster ensemble approach for categorical data clustering, and they improved the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble.

There are also some disadvantages when the number of base clusterings is large. For example, computing and storage overhead of system is greatly increased, and the difference between the base clusterings will continue to decrease. Zhou et al. [14] proposed the selective ensemble, and proved that the performance of the integration of some clustering results is better than the integration of all clustering results. Fern et al. [15] designed three different selection approaches of JC (Joint Criterion), CAS (Cluster and Select), CH (Convex Hull) that jointly consider quality and diversity. Azimi et al. [16] presented an adaptive cluster ensemble selection, and they proposed a novel framework that selects ensemble members for each data set based on its own characteristics. Jia et al. [17] developed a similarity-based spectral clustering ensemble selection, and they used the random scaling parameter, Nystr$\ddot{o}$m approximation and random initialization of k-means to perturb spectral clustering for producing the components of an ensemble system. Liu et al. [18] advanced a new selective clustering ensemble algorithm, they used the compactness and the separation to measure the quality of the clustering and defined the connectivity matrix to measure the quality and diversity.

We propose a new method based on covariance to measure the diversity. Firstly, base clustering results are generated by K-Means, AP, and FCM. Secondly, we calculate the covariance between each of the two base clustering results, and generate covariance matrix. Finally, part of base clustering results with small covariance are chosen to ensemble by CSPA.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces the principle of selective cluster ensemble based on covariance. Section 4 reports the experimental results. Section 5 provides conclusions and future work.

2 Related Work

It is a key point to measure the diversity of clustering results in selective clustering ensemble. Fern [19] used the normalized mutual information (NMI) to measure the diversity of clustering results.

$$NMI = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{1}$$

where $I(X,Y)$ is the mutual information of random variable X and Y, $I(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$, $H(X)$ is the he entropy of the X, $H(Y)$ is the entropy of the Y, and $H(X) = \sum_x p(x) log \frac{1}{p(x)}$. The NMI value is between 0 and 1, the value is smaller, the diversity is lager. Unlike other measure methods, NMI is not biased by large clusters.

Derek [20] used a method based on entropy to measure the diversity of clustering results.

$$div(c) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} -(p_{ij} log_2 p_{ij} + (1-p_{ij}) log_2 (1-p_{ij})) \qquad (2)$$

where p_{ij} is the probability of x_i and x_j are cluster in the same class, $p(x,y) = \frac{1}{k} \sum_{h=1}^{k} \delta(\pi_h(x_i), \pi_h(x_j))$, $\pi_h(x_i)$ is the label of the x_i in the class π_h, and $\pi_h(x_j)$ is the label of the x_j in the class π_h. If $\pi_h(x_i) = \pi_h(x_j)$, δ is 1, otherwise δ is 0. The value is also between 0 and 1, the value is smaller, the diversity is smaller.

Hadjitodorov [21] proposed four methods based on the adjusted rand index to measure the diversity of clustering results, and discovered the performance of the ensemble by middle value of diversity is better than the ensemble by max value of diversity.

$$ar(\pi_a, \pi_b) = \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} \binom{n_{h,l}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \qquad (3)$$

where $t_1 = \sum_{h=1}^{k_a} \binom{n_h}{2}$, $t_2 = \sum_{l=1}^{k_b} \binom{n_l}{2}$, $t_3 = \frac{2t_1 t_2}{N(N-1)}$, k_a and k_b are the number of clusters of π_a and π_b, respectively, $n_{h,l}$ is the number of points that are the same time in the cluster h and the cluster l, n_l is the number of points in the cluster l, and n_h is the number of points in the cluster h. The value is smaller, the diversity is lager. When two clusters are completely independent, the value is 0.

Luo [22] proposed five methods to measure the diversity, including CEBDM based on conditional entropy, DFBDM based on double fault measure, CFDBDM based on coincident failure diversity and IRABDM based on measurement of inter-rater agreement. The values of five methods are smaller, the diversity is smaller. Li [23] proposed a new method based on support vector machine to measure the diversity. Zhou [24] described in details some other methods of pairwise measures and non-pairwise measures, including Q-Statistic, Kohavi-Wolpert variance and so on.

3 Selective Clustering Ensemble Based on Covariance

We propose a new method based on covariance to measure the diversity. Covariance is a method used to measure the correlation between random variables, and the clustering result is deemed to the random variable, so the covariance is used to measure the diversity of clustering results. In addition, unlike NMI and

CE also consider expectation and variance after obtaining values, the covariance uses the expectation in calculating the value, so the covariance has been considered the problem of the offset. Let (X, Y) be a two-dimensional random variable, $E(X)$ and $E(Y)$ were the expectation of X and Y, respectively. $COV(X, Y)$ is the covariance between X and Y, as follows,

$$COV(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y). \quad (4)$$

Let $\pi(x_i)$ be the label of the x_i, $\pi(x_j)$ be the label of the x_j. We define a formula as follows,

$$\pi(x_i) - \pi(x_j) = \begin{cases} 1 & \pi(x_i) \neq \pi(x_j) \\ 0 & \pi(x_i) = \pi(x_j) \end{cases} . \quad (5)$$

For an n-dimensional random variable $X = (X_1, X_2, ..., X_n)$, let $\sigma_{ij} = COV(X_i, X_j)$, $i, j = 1, 2, ..., n$, it defines matrix V is the covariance matrix of X, and V is an n-order symmetric matrix.

$$V = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix} \quad (6)$$

where $\sigma_{11} = COV(X_1, X_1)$ is the variance of X_1.

N clustering results are deemed to an n-dimensional random variable $X = (X_1, X_2, ..., X_n)$. The covariance matrix V is a symmetric matrix, and the values on the diagonal are variance. We only consider the difference between the base clustering results, and don't consider the positive correlation and negative correlation, so we simplify V to V' that all values are non-negative and values on the diagonal are 0. And it is

$$V = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix} \longrightarrow V' = \begin{bmatrix} 0 & \sigma_{12} & \cdots & \sigma_{1n-1} & \sigma_{1n} \\ 0 & 0 & \cdots & \sigma_{2n-1} & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \sigma_{(n-1)n} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} . \quad (7)$$

The steps of select base clustering results to ensemble(SBCRE) are shown in Algorithm 1. The input is m base clustering results, and the output is m ensemble results. Firstly, we calculate the covariance between two base clustering results, and generate covariance matrix V. Secondly, V is simplified to V'. Thirdly, we select the maximum value of V', remove the row with maximum, and set the maximum to 0. Fourth, the ensemble results are obtain by ensemble the remaining base clustering results with CSPA. Finally, the output is obtained with m iterations.

Algorithm 1. SBCRE

```
Input: $m$ base clustering results
Output: $m$ ensemble results
begin
    Calculate the covariance between two base clustering results according
    to the formula (4), and generate covariance matrix $V$;
    Simplify $V$ to $V'$ according to the formula (7);
    if $m \geqslant 1$ then
        Select the maximum value of $V'$, record row number $r$ and column
        number $c$;
        Remove the $rth$ row;
        The maximum is set to 0, update $V'$;
        $m = m - 1$;
        The ensemble results are obtain by ensemble the $m$ base clustering
        results with CSPA.
    end
end
```

We uses three different cluster methods of K-Means [25,26], AP [27] and FCM [28,29]. The 60 base clustering results are generated with different initialization. We get 60 ensemble results by Algorithm 1, and calculate the F-measure between each ensemble result and the label of each data set. The final result with maximum F-measure on each data set is obtained. The steps of selective clustering ensemble based on covariance(SCEBC) are shown in Algorithm 2.

Algorithm 2. SCEBC

```
Input: The data set $X$ has $n$ samples
Output: The set has labels of $n$ samples
begin
    Generate 20 base clustering results according to the K-Means;
    Generate 20 base clustering results according to the AP;
    Generate 20 base clustering results according to the FCM;
    Get $m$ ensemble results according to the Algorithm 1;
    Calculate the F-measure between each ensemble result and the label of data set;
    The ensemble result with the maximum F-measure as the output.
end
```

4 Experiment

4.1 Data Set

The 20 UCI data sets are used in the experiment. The number of features, classes and instances on each data set are shown in Table 1.

Table 1. The number of features, classes and instances on each data set

Data Set	Features	Classes	Instances
Iris	4	3	150
Glass	9	6	214
Wine	13	3	178
Zoo	16	7	101
Ionosphere	34	2	351
Sonar	60	2	208
Balance scale	4	3	625
Pima	8	2	768
Spect-heart	22	2	267
Hepatitis	19	2	155
Bupa	6	2	345
Habermans survival	3	2	306
Wdbc	30	2	569
Statlog	19	7	2310
Vehicle	18	4	846
Breast-cancer-Wisconsin	9	2	683
Car	6	4	1728
Credit-g	20	2	1000
Vowel	13	11	990
Lymphography	18	4	148

4.2 Evaluation Criteria

F-measure is the evaluation criteria of experiment results [30], and it is shown in formula (8).

$$F(i) = \frac{2 \times precision(i,j) \times recall(i,j)}{precision(i,j) + recall(i,j)} \tag{8}$$

where $precison(i,j) = \frac{N_{ij}}{N_i}$ is the precision, $recall(i,j) = \frac{N_{ij}}{N_j}$ is the recall, N_i is the total number of samples of correct clustering, N_j is the total number of samples of jth class in clustering results, and N_{ij} is the total number of correct clustering of jth class in clustering results. However, the formula (8) will get a lot of $F(i)$ values, so the F-measure is weighted and averaged by formula (9), as follows,

$$F(i)' = \frac{\sum_{i=1}^{k}(|i| \times F(i))}{\sum_{i=1}^{k}|i|}. \tag{9}$$

4.3 Experiment Result

The experiment results are reported in this subsection. The F-measures of different algorithms on each data set are shown in Table 2, where ALL is directly ensemble, RSE is selective ensemble based on random, and CSEV is average

Table 2. The F-measures of different algorithms on the each data set

Data Set	Base Clustering Methods			Clustering Ensemble by CSPA		
	K-means	AP	FCM	ALL	RSE	CSEV
Iris	0.8519	0.8351	0.8644	0.8667	0.8783	**0.8812**
Wine	0.6598	0.6618	0.6622	0.6632	0.6636	**0.6689**
Zoo	0.6010	0.6183	0.6273	0.6319	0.6223	**0.6329**
Glass	0.5581	0.5268	0.5281	0.5390	0.5252	**0.5867**
Ionosphere	0.6708	0.6670	0.6792	0.6815	0.6650	**0.6902**
Sonar	0.5428	0.4636	0.5425	0.5476	0.5389	**0.5506**
Balance scale	0.5511	0.5410	0.5554	0.5629	0.5579	**0.5672**
Pima	0.5547	0.5469	0.5433	0.5656	0.5657	**0.5658**
Spect-heart	0.6112	0.6114	0.6096	0.6188	0.6169	**0.6234**
Hepatitis	0.5335	0.5232	0.5266	0.5409	0.5364	**0.5519**
Bupa	0.6429	0.5538	0.6080	0.6674	0.6623	**0.6687**
Habermans survival	0.5951	0.5934	0.5948	0.5952	**0.5953**	**0.5953**
Wdbc	0.8639	0.8758	0.8639	**0.8850**	0.8747	**0.8850**
Statlog	0.5478	0.5490	0.5456	0.5477	0.5473	**0.5495**
Vehicle	0.4691	0.4580	0.4632	0.4747	0.4749	**0.4757**
Breast-cancer-Wisconsin	0.7606	0.7592	0.7564	0.7963	0.7969	**0.7997**
Car	0.3253	0.3219	**0.4817**	0.3186	0.3209	0.3476
Credit-g	0.5009	0.4952	0.4976	0.5185	0.5029	**0.5385**
Vowel	0.1966	0.1973	0.1649	0.1992	0.1944	**0.2004**
Lymphography	0.5029	0.4820	0.4866	0.4905	0.4894	**0.5085**

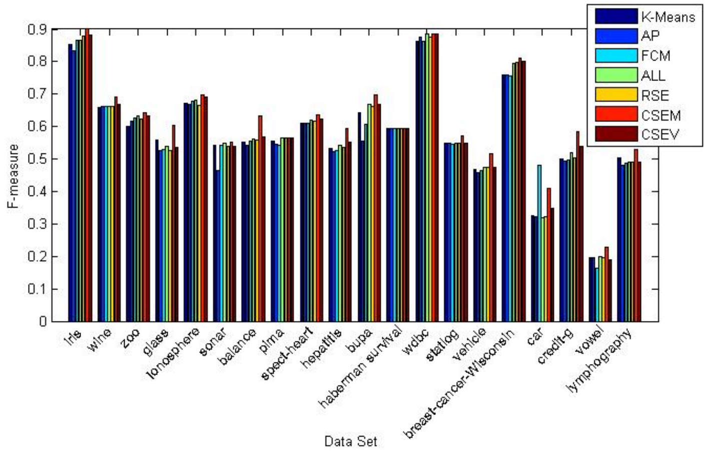

Fig. 1. The F-measure of different algorithms on the each data set

value of selective ensemble based on covariance. A F-measure value between an ensemble result and the labels of data set is obtained with one iteration, so a total of 60 F-measure values are obtained. The CSEV is the average value of the 60 F-measure values.

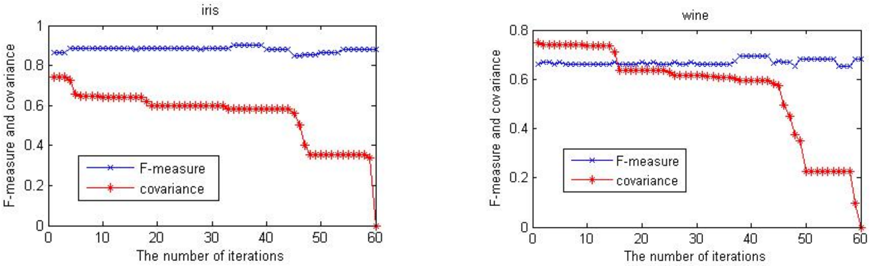

Fig. 2. The F-measure and covariance between each ensemble result and the label of Iris and Wine

Table 3. The covariances with maximum F-measure on each data set

Data Set	Covariance	maximum F-measure
Iris	0.5803	0.9000
Wine	0.5940	0.6924
Zoo	0.5822	0.6438
Glass	0.5029	0.6045
Ionosphere	0.2498	0.6977
Sonar	0.4236	0.5524
Balance scale	0.4416	0.6326
Pima	0.2360	0.5658
Spect-heart	0.2474	0.6352
Hepatitis	0.1425	0.5927
Bupa	0.1817	0.6990
Habermans survival	0.3827	0.5953
Wdbc	0.2493	0.8870
Statlog	0.1846	0.5698
Vehicle	0.4248	0.5152
Breast-cancer-Wisconsin	0.2229	0.8113
Car	0.2135	0.4109
Credit-g	0.1489	0.5834
Vowel	0.5993	0.2268
Lymphography	0.2950	0.5285

From the Table 2, we can see that the F-measures of clustering ensemble are better than base clustering on 16 data sets except Glass, Statlog, Car, and Lymphography. The F-measure of CSEV equals RSE on the Habermans survival, the F-measure of CSEV equals ALL on the Wdbc, the F-measure of FCM is better than CSEV on the Car, and the F-measures of CSEV are better than base clustering, ALL, and RSE on other 17 data sets.

We can obtain two conclusions based on above results. Firstly, the clustering ensemble result is better than base clustering. Secondly, the CSEV is better than base clustering, ALL, and RSE, which can also be seen from Fig. 1, where CSEM is max value of selective ensemble based on covariance. The x axis of Fig. 1 are 20 data sets and the y axis are the F-measures.

The covariances with maximum F-measure on each data set are shown in Table 3. From the Table 3, we can see that the covariance is between 0.1 and 0.6 on 20 data sets. Therefore, we will be directly select base clustering results that covariance in this interval to ensemble in the practical applications.

We can clearly see all F-measures and covariances of each selection on iris, wine, zoo, and glass from Fig. 2 to Fig. 3. The x axis is the number of base clustering results that does not use to ensemble. The y axis are F-measures and covariances between each ensemble result and the label of each data set.

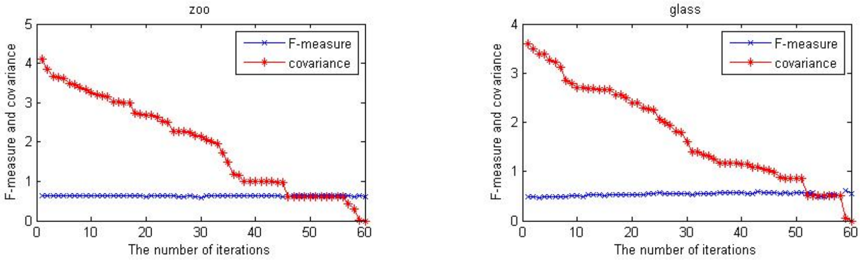

Fig. 3. The F-measure and covariance between each ensemble result and the label of Zoo and Glass

5 Conclusion

In this paper, we propose the selective clustering ensemble based on covariance. We measure diversity based on covariance. Our work may prove the better performance of our algorithm with experiments on the 20 UCI data sets, and get a covariance interval that is between 0.1 and 0.6. In future work, we will try to study more on selective clustering ensemble based on covariance and use them to the practical applications. We also will try to add semi-supervised information in this algorithm and achieve the parallelization of this algorithm.

Acknowledgements. This work is supported by the National Science Foundation of China (Nos. 61170111, 61003142, 611734002), the Fundamental Research Funds for the Central Universities(No. SWJTU11ZT08) and the Research Fund of Traction Power State Key Laboratory, Southwest Jiaotong University (No. 2012TPL_T15).

References

1. Han, J.W., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2007)
2. Tang, W., Zhou, Z.H.: Selective Clustering Ensemble Based on Bagging. Journal of Software 16(4), 496–502 (2005)
3. Strehl, A., Ghosh, J., Cardie, C.: Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research (3), 583–617 (2002)
4. Topchy, A., Jain, A.K., Punch, W.: A Mixture Model for Clustering Ensembles. In: Proc. of the 4th SIAM International Conference on Data Mining, pp. 379–390 (2004)
5. Topchy, A., Jain, A.K., Punch, W.: Clustering Ensembles: Models of Consensus and Weak Partitions. IEEE Trans. on Pattern Analysis and Machine Intelligence 21(12), 1866–1881 (2005)
6. Luo, H.L., Jing, F.R., Xie, X.B.: Combining Multiple Clusterings Using Information Theory Based Genetic Slgorithm. In: IEEE International Conference on Computational Intelligence and Security, vol. 1, pp. 84–89 (2006)
7. Hassan, S.Z., Verma, B.: Decisions Fusion Strategy: Towards Hybrid Cluster Ensemble. In: Intelligent Sensors, Sensor Networks and Information, pp. 377–382 (2007)
8. Mohammadi, M., Nikanjam, A., Rahmani, A.: An Evolutionary Approach to Clustering Ensemble. In: The Fourth International Conference on Natural Computation, vol. 3, pp. 77–82 (2008)
9. Iqbal, A.M., Moh'd, A., Khan, Z.A.: Semi-supervised Clustering Ensemble by Voting. In: The International Conference on Information and Communication System, pp. 1–5 (2009)
10. Wang, H.J., Li, Z.S., Qi, J.H.: Semi-Supervised Cluster Ensemble Model Based on Bayesian Network. Journal of Software 21(11), 2814–2825 (2010)
11. Jia, J.H., Xiao, X., Liu, B.X.: Bagging-Based Spectral Clustering Ensemble Selection. Pattern Recognition Letters 32(10), 1456–1467 (2011)
12. Yang, Y., Wang, H., Lin, C., Zhang, J.: Semi-supervised Clustering Ensemble Based on Multi-ant Colonies Algorithm. In: Li, T., Nguyen, H.S., Wang, G., Grzymala-Busse, J., Janicki, R., Hassanien, A.E., Yu, H. (eds.) RSKT 2012. LNCS (LNAI), vol. 7414, pp. 302–309. Springer, Heidelberg (2012)
13. Iam-On, N., Boongoen, T., Garrett, S., Price, C.: A Link-Based Cluster Ensemble Approach for Categorical Data Clustering. IEEE Transactions on Knowledge and Data Engineering 24(3), 413–425 (2012)
14. Zhou, Z.H., Wu, J., Tang, W.: Ensembling Neural Networks: Many Could be Better Than All. Artificial Intelligence 137(1-2), 239–263 (2002)
15. Fern, X.Z., Lin, W.: Cluster ensemble selection. Statistical Analysis and Data Mining 1(3), 128–141 (2008)
16. Azimi, J., Fern, X.Z.: Adaptive cluster ensemble selection. In: Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI 2009), pp. 992–997 (2009)
17. Jia, J.H., Xiao, X., Liu, B.X.: Similarity-based Spectral Clustering Ensemble Selection. In: Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), pp. 1071–1074 (2012)
18. Liu, L.M., Fan, X.P.: A New Selective Clustering Ensemble Algorithm. In: Proceedings of the 9th International Conference on e-Business Engineering (ICEBE), pp. 45–49 (2012)

19. Fern, X.Z., Brodley, C.E.: Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003), pp. 186–193 (2003)
20. Derek, G., Alexey, T., Nadia, B.: Ensemble Clustering in Medical Diagnostics. In: Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, CBMS, pp. 576–581 (2004)
21. Hadjitodorov, S.T., Kuncheve, L.I., Todorova, L.P.: Moderate Diversity for Better Cluster Ensembles. Information Fusion 7(3), 264–275 (2006)
22. Luo, H.L., Kong, F.S., Li, Y.X.: Diversity Measure of Cluster Ensemble. Journal of Computers 30(8), 1315–1324 (2007)
23. Li, K., Gao, H.T.: A Novel Measure of Diversity for Support Vector Machine Ensemble. In: Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 367–370 (2010)
24. Zhou, Z.H.: Ensemble Methods: Foundations and Algorithms, Boca Raton, FL (2012)
25. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
26. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28(1), 100–108 (1979)
27. Frey, J.B., Dueck, D.: Clustering by Passing Messages Between Data Points. Science 315(5814), 972–976 (2007)
28. Dunn, J.C.: A Graph Theoretic Analysis of Pattern Classification via Tamura's Fuzzy Relation. Journal of Cybernetics 3(3), 32–57 (1973)
29. Bezdek, J.C.: A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 2(1), 1–8 (1980)
30. Yang, Y., Kamel, M.: An Aggregated Clustering Approach Using Multi-Ant Colonies Algorithms. Pattern Recognition 39(7), 1278–1289 (2006)