A Clustering Based Feature Selection Method Using Feature Information Distance for Text Data

Shilong Chao, Jie Cai, Sheng Yang(✉), and Shulin Wang

College of Computer Science and Electronic Engineering,
Hunan University, Changsha, China
Yangsh0506@sina.com

Abstract. Feature selection is a key point in text classification. In this paper a new feature selection method based on feature clustering using information distance is put forward. This method using information distance measure builds a feature clusters space. Firstly, K-medoids clustering algorithm is employed to gather the features into *k* clusters. Secondly the feature which has the largest mutual information with class is selected from each cluster to make up a feature subset. Finally, choose target number features according to the mRMR algorithm from the selected subset. This algorithm fully considers the diversity between features. Unlike the incremental search algorithm mRMR, it avoids prematurely falling into local optimum. Experimental results show that the features selected by the proposed algorithm can gain better classification accuracy.

Keywords: Text classification · Feature selection · Cluster · Diversity

1 Introduction

Text classification is one of the important branches of data mining. Its target is to classify the text sets according to a certain standard or system automatically [1, 2]. In text data, Vector Space Model (VSM) [3–5] is widely used to represent the text. The VSM of text data often has high dimensionality. In order to exactly classify the text, it is necessary to reduce the dimensionality of text data by feature selection [6]. Feature selection can remove the irrelevant features, at the same time redundant features also are effectively eliminated. Finally the feature subset composed of features with strong distinguishing ability can be selected, and the accuracy and speed of classification are improved.

Feature selection based on information theory has always been a hot research topic, and a number of classical mutual information (MI) based algorithms have been proposed. Battiti proposed an algorithm called MIFS [7] which takes into account both feature-feature and feature-class mutual information for feature selection. MIFS works well when the penalty parameter is set appropriately. Peng and Ding put forward the classic mRMR algorithm [8]. The idea of mRMR is to maximize the correlation between feature and class, while minimizing the redundancy between the selected

D.-S. Huang et al. (Eds.): ICIC 2016, Part I, LNCS 9771, pp. 122–132, 2016.
DOI: 10.1007/978-3-319-42291-6_12

features. In mRMR, the multivariate joint probability density estimation problem of high dimensional space is replaced by the probability densities of couples. FCBF algorithm [9] uses the symmetrical uncertainty (SU) to measure the correlations of feature-class and feature-feature. In the original feature set, the features whose SU values with the class are less than the given threshold are deleted, and then the redundancy analysis is carried out in the remaining features. The final feature subset is obtained after the redundant features are eliminated. The CMIM algorithm [10] proposed by Fleuret takes the conditional mutual information as the measure to select features. Fleuret thinks that the more conditional mutual information the feature has in the case of selected subset, the more information about class the feature carries. The idea of CMIM is to select the feature with maximal conditional mutual information in the case of selected subset. Literature [11] made a more comprehensive and detailed summary of the feature selection based on mutual information.

Information measurements can be used to measure the uncertainty and the non-linear relationship between the features in quantitative form [12, 13]. As a result the information measures are widely used in feature selection. These above algorithms employ relevant metrics to measure the relationships of feature-class and feature-feature. Their common basic idea is to select those features relevant with class and irrelevant between selected features. These algorithms take advantage of the greedy algorithm which is easy to reach a local optimum. From the overall situation, clustering analysis on the original feature set can be considered. If the clustering analysis is applied to feature selection, the prematurely local optimal solution may be avoided. Firstly, clustering is performed on the original feature set. The features with high correlation and high redundancy are clustered together, and features in different clusters have larger diversity. The feature strongly associated with the class is selected into a feature subset from each cluster. In this way, the redundant degree of the selected feature subset is relatively low, and the correlation with class is strong. The process of feature selection based on feature clustering almost starts with calculating the corresponding measure of the specified data set. Features are clustered according to specified clustering algorithm. At last the representative features of each cluster are selected to compose the final feature subset.

Au and Chan proposed the ACA [14] algorithm, it chooses the information measurement R to measure the correlation between the features. They use K-means to cluster the features, and then select the representative feature of each class. The algorithm is very perfect in the classification of gene data. The disadvantage of this algorithm is that sometimes the clustering may enter a dead cycle, and it is needed to specify a certain number of iterations to terminate the clustering process. Song developed an algorithm FAST [15] which uses hierarchical clustering method. In this algorithm, many minimum spanning trees are built with the measure SU and each tree is treated as a class cluster. FAST is suitable for high dimensional data. The disadvantage is that the number of selected features is determined by the data. Liu also introduced a feature clustering feature selection algorithm MFC [16] based on minimum spanning tree. Different from FAST the measure of MFC is Variation of Information (VI). Obviously, using the diversity between features in feature selection has already become a hot spot.

2 Information Distance Measure Based Feature Clustering

2.1 Basic Idea

The algorithm proposed in this paper mainly considers the diversity between the features. The diversity metric namely the information distance is used to measure the redundancy of the feature subset. The greater diversity the feature subset has, the lower redundancy it has.

Information distance [17–19] is a kind of diversity measurement based on information theory. In the literature [17], the paper listed the most commonly used measures based on information theory. Here we choose the measure D:

$$\begin{aligned} D &= \frac{1}{2}[H(X)+H(Y)] - I(X;Y) \\ &= \frac{1}{2}[H(X|Y)+H(Y|X)] \\ &= \frac{1}{2}(\sum_{x_i}\sum_{x_j} p(x_i x_j)\log\frac{1}{p(x_i|x_j)} + \sum_{x_i}\sum_{x_j} p(x_i x_j)\log\frac{1}{p(x_j|x_i)}) \\ &= \frac{1}{2}(\sum_{x_i}\sum_{x_j} p(x_i x_j)\log\frac{1}{p(x_i|x_j)p(x_j|x_i)}) \end{aligned} \tag{1}$$

In formula (1), X and Y are random variables, $H(X)$ and $H(Y)$ are the information entropies of X and Y, and $I(X;Y)$ is the mutual information between X and Y. The conditional entropy $H(X|Y)$ represents the conditional uncertainty of X given Y. $p(x_i x_j)$ is the joint probability density between x_i and x_j. $p(x_i|x_j)$ is the conditional probability density of x_i given x_j.

The standard distance measure must satisfy the three properties: non-negativity, symmetry and triangle inequality. It had been proved that the information distance D is in line with the three properties in the literature [17].

$$D(X,Y) \geq 0(\text{non} - \text{negativity}) \tag{2}$$

$$D(X,Y) = D(Y,X)(\text{symmetry}) \tag{3}$$

$$D(X,Y) \leq D(X,Z) + D(Y,Z)(\text{triangle inequality}) \tag{4}$$

Essentially, it has been recognized that feature selection needs to select a subset with strong discriminatory power. The features have high relevance with class should be chosen. However, experiments had proved that the m best features are not the best m features [20, 21]. There is redundancy among these features. Therefore, the selected subset should achieve a balance point between the relevance about class and the redundancy among features. If a feature subset S has been selected, the relevance between S and class is *Rel*, and the redundancy among S is *Red*.

$$Rel = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) \tag{5}$$

$$Red = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i; x_j) \tag{6}$$

$$MRMR : \max(Rel - Red) \tag{7}$$

In order to reach the balance, on the one hand the selected subset of features should guarantee the most relevance with class, while on the other hand the degree of redundancy among the feature subset must be lowest. The criterion combing the formulas (5) and (6) is MRMR. In practice, incremental search algorithm mRMR was used to find the near-optimal feature subset. Assuming that the original feature set is F, the subset S_{m-1} of the m-1 features is selected. Now the target should be to choose the mth feature from the feature subset $\{F - S_{m-1}\}$ which satisfies formula (8). Here, mRMR is employed to select features from a refined feature subset that produced by feature clustering using information distance measure.

$$\max_{f_j \in F - S_{m-1}} [I(f_j; c) - \frac{1}{m-1}\sum_{f_i \in S_{m-1}} I(f_j; f_i)] \tag{8}$$

2.2 K-medoids Based Feature Clustering

Clustering is a commonly used unsupervised learning method of data analysis. It can divide the data according to the diversity between each other. Similarly, cluster analysis can also be applied to the features. Actually feature clustering has been widely used in feature selection methods.

K-means and K-medoids are two classic unsupervised clustering algorithms based on partitioning. According to distance between the objects in data set, K-means could divide the objects into several clusters. The distance measure generally use geometric distance such as Minkowski distance, Euclidean distance, Manhattan distance. Suppose the Euclidean distance is chosen as the distance measure. After deciding the k center randomly, the non-center points of data set are divided to the nearest center point. Clusters are formed after every object has been assigned. The average geometric coordinate value of each cluster is set to be the new center point of the cluster. Repeat the process until the clusters are stable. However, Features are a set of discrete points in the feature information distance space, and the features have no coordinate. Therefore, K-means algorithm can't be used for feature clustering, and K-medoids clustering algorithm is adopted for feature clustering. The measure employed is the information distance D mentioned above. In the iteration process, the values of the D between the features are constant and just need calculate once, which greatly reduces the computational complexity of the algorithm.

Unlike K-means, K-medoids algorithm selects actual object as the cluster center, rather than using the mean as the center in the cluster. K-medoids clustering algorithm is based on the principle that minimizes the degree of diversity between cluster center and all the objects in data set. Corresponding to the feature cluster, the information distance sum within the cluster should be smallest. An error criterion is designed in formulas (9) and (10) to measure the cost of replacing the original center by any of the non-center features in the cluster.

$$E = \sum_{i=1}^{k} \sum_{f \in C_i} D(f, o_i) \tag{9}$$

$$T = E_f - E_{o_i} \tag{10}$$

E is the sum of the information distance between all the features and corresponding center in the data set. $D(f, o_i)$ is the information distance between feature f and nearest center o_i. T is the cost of the cluster center o_i replaced by a non-center feature f. At the initial stage of the K-medoids k features are selected randomly as the center, and then the remaining features are assigned to the nearest center in information distance. In the iterations, for each non-center feature calculate the cost T of replacing the center in each cluster. If the cost T is negative, the actual error E must be reduced, and the current center feature can be replaced by this feature. Here the original center should be replaced by the one with the minimal T in which the convergence speed can be accelerated. Conversely the center won't be changed. If all the centers don't change any more, clustering process is over.

3 Algorithm and Steps

According to the above mentioned, we propose a text data feature selection method called Information Distance Measure based Clustering for Feature Selection (IDMCFS). Its process framework is described as Fig. 1.

IDMCFS is divided into two stages: clustering stage and selecting stage. The clustering stage: assuming that the size of ultimate target features is m, the original feature set is clustered into k clusters using K-medoids. k is much larger than m. From each cluster select the feature whose mutual information with the class is largest to form the candidate feature subset S'. In the selecting stage, m features are selected from subset S' according to mRMR. The redundancy of selected features are further reduced and the diversity between the features are guaranteed. The final m features is the solution of IDMCFS. Algorithm is described as follows.

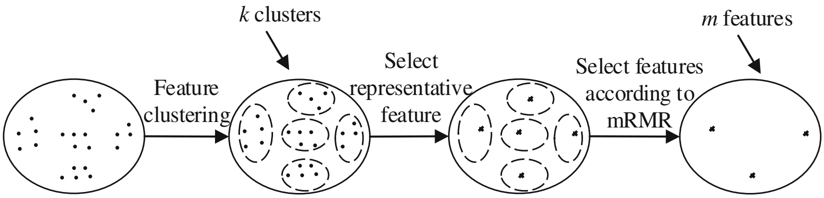

k clusters
m features
Feature clustering
Select representative feature
Select features according to mRMR

Fig. 1. The process framework of IDMCFS

```
Input:  $F = \{f_1, f_2, ..., f_n\}$ : original feature set
        $m$:target feature number
        $k$: cluster number, $k > m$
Output: $S$: selected feature subset
```

```
Begin
   $Center := \Phi$ , Subset $S' := \Phi$ , $S := \Phi$ ;
  For each $f_i \in F, f_j \in F$
     calculate $H(f_i)$ , $I(f_i, c)$ and $D_{ij}$ ;
  Repeat
     Randomly select a $f \in F$ ;
     $Center := Center + f$ ;
     $F := F - f$ ;
  Until $|Center| \geq k$
  Repeat
     For each $f \in F$
        Divide $f$ to nearest center;
     For each $f \in F$
        Compute $T_f$ according to formula(10);
     For each cluster
        Select $f$ in cluster with minimal $T_f$ to replace center;
  Until Center not change
  Repeat
     Select $f$ in each cluster with maximal $I(f, c)$ ;
     $S' := S' + f$ ;
  Until all clusters have been processed
  Select $f \in S'$ with maximal $I(f, c)$ ;
  $S := S + f$ ;
  Repeat
     Choose $f \in S'$ according to formula(8);
     $S := S + f$ ;

     $S' := S' - f$ ;
  Until $|S| \geq m$
  Return $S$;
end
```

The complexity of each iteration in the clustering phase of IDMCFS algorithm is $O\left(k(n-k)^2\right)$ (k is the number of clusters and n is the number of features in data set). Assuming that the total times of iterations is l, and then clustering complexity is $O\left(lk(n-k)^2\right)$. Because the feature with minimal T value is selected as the new center in iteration, thus the convergence speed is greatly accelerated and the number of iteration is very small. The complexity of selecting representative features and running mRMR are $O(n)$ and $O(km)$ respectively. Overall speaking, the complexity of IDMCFS algorithm is $O\left(lk(n-k)^2\right)$. When the cluster number k is much smaller than the original feature number n, the complexity will be $O(n^2)$.

4 Experiment Analysis

The experimental data sets are derived from the open data set (http://www.tunedit.org/repo/Data/Text-wc). These data sets are designed to test the classification performance of text data. In order to facilitate the calculation of mutual information, the data are discretized by MDL [22] discretization method. The information of data sets after discretized is displayed in Table 1. All the experiments are carried out on the experimental platform of Matlab2014a and Weka3.7. Weka is used to discretize the data and gain the classification accuracy rate of the final feature subset. The main procedure of IDMCFS is realized by Matlab2014a.

Table 1. Information of datasets

Dataset	Features	Instances	Classes
tr12.wc	126	313	8
tr11.wc	250	414	9
wap.wc	545	1560	20
la1s.wc	1438	3204	6
la2s.wc	1438	3075	6
fbis1.wc	1138	2463	17

In order to eliminate the influence of generating initial center randomly in K-medoids algorithm to the clustering results, the experiment was repeated 50 times and took the average value as the final result. The performance of IDMCFS was compared to mRMR, CMIM and ReliefF algorithms on the same data sets. To check the performance of selected feature subset, the Naive Bayes classifier and 10-fold cross validation were adopted to test the accuracy rates of classification.

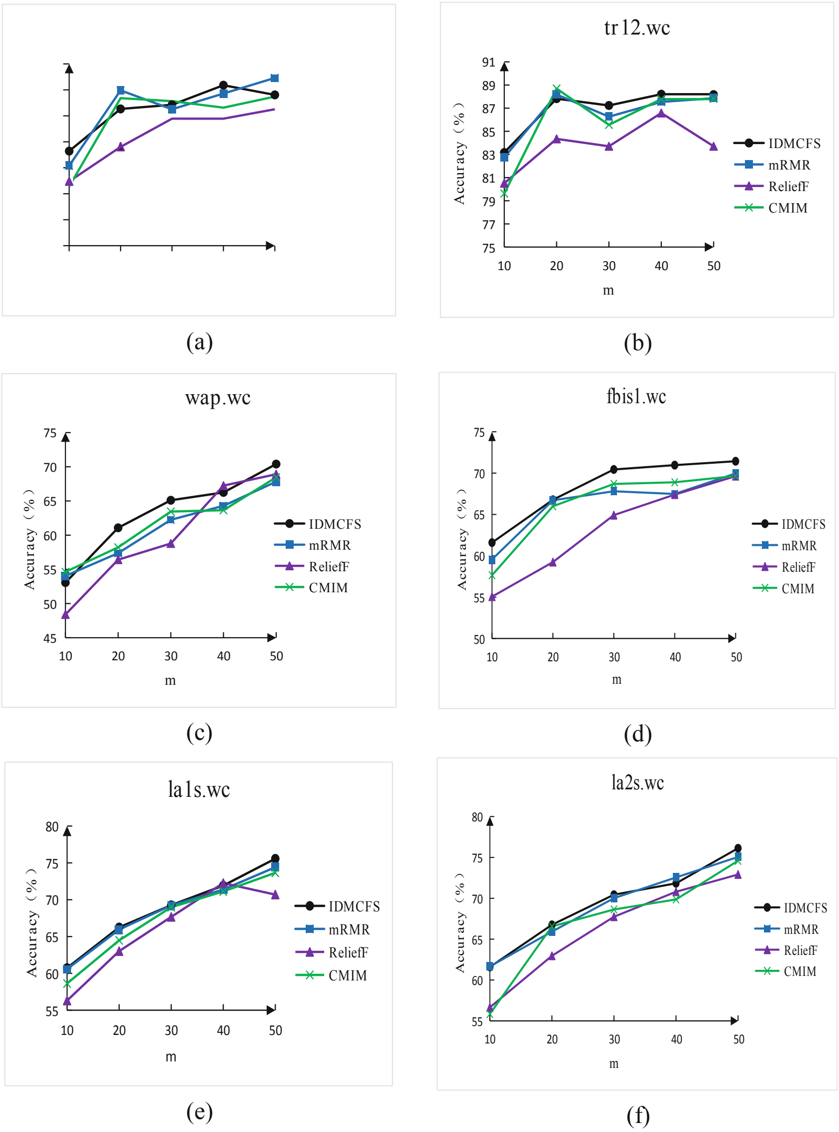
Table 2 shows the analogy of IDMCFS algorithm in different data sets and different parameters. It can be seen that in each data set there is an identical tendency. With the number of selected features m growth, the classification accuracy rates have increased. As we all know, more features often have greater distinguishing ability. Moreover, if the number of target features m is determined, the accuracy rates of classification don't

Table 2. Classification accuracy (%) of IDMCFS in different situations (*m* number of selected features, *k* number of clusters)

m	*k*	tr11.wc	tr12.wc	wap.wc	la1s.wc	la2s.wc	fbis1.wc
10	20	80.19	79.36	49.85	57.18	58.86	59.31
	30	81.37	**83.28**	51.36	58.65	60.35	60.09
	40	**82.56**	83.18	52.20	**60.74**	61.58	61.59
	50	82.29	83.16	**53.06**	60.46	**61.59**	**61.66**
20	30	82.31	84.02	58.68	63.97	65.84	65.63
	40	83.38	86.91	**61.07**	64.93	66.16	66.22
	50	84.23	87.82	59.84	65.66	**66.76**	**66.78**
	60	**85.53**	**88.12**	60.31	**66.25**	65.96	66.15
30	40	83.67	85.43	64.41	68.78	69.98	69.76
	50	85.24	86.55	65.01	68.80	70.21	69.89
	60	**85.85**	**87.23**	**65.10**	69.00	70.39	70.22
	70	85.58	87.20	64.77	**69.26**	**70.44**	**70.44**
40	50	85.26	86.18	63.52	70.44	70.51	68.89
	60	86.33	86.35	64.57	71.16	70.77	69.13
	70	**87.66**	**87.65**	65.69	**71.91**	70.94	69.97
	80	87.46	88.20	**66.25**	71.90	**71.83**	**70.97**
50	60	85.12	86.20	69.45	71.24	70.87	68.93
	70	85.55	87.96	**70.40**	72.64	72.56	69.54
	80	**86.61**	88.17	70.15	72.99	73.70	70.56
	90	86.16	**88.37**	70.09	**75.57**	**76.12**	**71.44**

increase as the increasement of cluster number *k*. The performance of IDMCFS is not related to the cluster number *k*. As a result it is more likely to get the global optimal solution.

Figure 2 represents the comparisons of the proposed algorithm IDMCFS and mRMR, ReliefF, CMIM algorithms performance in different data sets. Overall, the trend curves of four algorithms are similar. The classification accuracies are improved with the increase of the selected feature number. When *m* is determined, CMIM and mRMR have their own advantages in different data sets, and the results of ReliefF are slightly worse. The classification accuracies of IDMCFS on most data sets are higher than that of mRMR, CMIM and ReliefF. This shows that the IDMCFS algorithm can select a feature subset which has greater ability to distinguish the text. According to the theory of IDMCFS, if cluster number *k* equals to the feature number of the original data set IDMCFS becomes mRMR. However, the experimental results demonstrate that the performance of IDMCFS is better than mRMR and CMIM. It's believable that better feature subset can be selected by fully considering the diversity between the features. Of course, the results in the table are just the best results got in the experiment, as long as the parameter *k* set properly, the actual effect of IDMCFS may be further improved.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. The accuracies of IDMCFS, mRMR, CMIM and ReliefF on each data set

5 Conclusion

In this paper, a feature selection algorithm based on feature clustering using information distance has been proposed. This algorithm mainly emphasizes the diversity between features, using an information distance measure to cluster the features. Highly redundant features are clustered into a cluster, and the features of different clusters are low relevant, so it is possible to obtain the global optimal solution. IDMCFS algorithm uses the K-medoids clustering algorithm which only needs the distance between features. At the same time, the number of clusters *k* is much larger than target feature number *m*, which assures the high redundancy between features in each cluster. The clustering process converges quickly. The overall performance of IDMCFS is better than mRMR, CMIM and ReliefF.

There is a challenge about IDMCFS that how to choose the proper cluster number. In this paper *k* was set as several constants which are larger than *m*. Hence the value of *k* may be not optimal. How to find the optimal *k* value will be the research direction in the future.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant No. 61472467).

References

1. Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 721–735 (2009)
2. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. **18**, 79–112 (2000)
3. Chen, Z., Lü, K.: A preprocess algorithm of filtering irrelevant information based on the minimum class difference. Knowl.-Based Syst. **19**, 422–429 (2006)
4. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**, 1–47 (2002)
5. Song, F., Liu, S., Yang, J.: A comparative study on text representation schemes in text categorization. Pattern Anal. Appl. **8**, 199–209 (2005)
6. Fragoudis, D., Meretakis, D., Likothanassis, S.: Best terms: an efficient feature-selection algorithm for text categorization. Knowl. Inf. Syst. **8**, 16–33 (2005)
7. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**, 537–550 (1994)
8. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005)
9. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)
10. Fleuret, F.: Fast binary feature selection with conditional mutual information. J. Mach. Learn. Res. **5**, 1531–1555 (2004)
11. Vinh, N.X., Epps, J., Bailey, J.: Effective global approaches for mutual information based feature selection. In: International Conference on Knowledge Discovery and Data Mining, pp. 512–521. ACM (2014)
12. Forman, G.: An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. **3**, 1289–1305 (2003)
13. Liu, H., Liu, L., Zhang, H.: Feature selection using mutual information: an experimental study. In: Ho, T.-B., Zhou, Z.-H. (eds.) PRICAI 2008. LNCS (LNAI), vol. 5351, pp. 235–246. Springer, Heidelberg (2008)
14. Au, W.H., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **2**, 83–101 (2005)

15. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. Knowl. Data Eng. **25**, 1–14 (2013)
16. Liu, Q., Zhang, J., Xiao, J., Zhu, H., Zhao, Q.: A supervised feature selection algorithm through minimum spanning tree clustering. In: IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 264–271 (2014)
17. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
18. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: 26th AI Conference, pp. 1073–1080 (2009)
19. Vinh, N.X, Epps, J.: A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. In: 9th IEEE International Conference on Bioinformatics and BioEngineering, pp. 84–91 (2009)
20. Jain, A.K., Duin, R.P., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 4–37 (2000)
21. Herman, G., Zhang, B., Wang, Y., Ye, G., Chen, F.: Mutual information-based method for selecting informative feature sets. Pattern Recogn. **46**, 3315–3327 (2013)
22. Fayyad, U., Irani, K.B.: Multi-interval discretization of continuous valued attributes for classification learning. In: 13th IJCAI, pp. 1022–1027 (1993)