AN ENSEMBLE CLUSTERING FOR MINING HIGH-DIMENSIONAL BIOLOGICAL BIG DATA

DEWAN MD. FARID, ANN NOWE & BERNARD MANDERICK
Computational Modeling Lab, Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2,
1050 Brussels, Belgium

ABSTRACT

Clustering of high-dimensional biological big data is incredibly difficult and challenging task, as the data space is often too big and too messy. The conventional clustering methods can be inefficient and ineffective on high-dimensional biological big data, because traditional distance measures may be dominated by the noise in many dimensions. An additional challenge in biological big data is that we need to find not only the clusters of instances (genes), but also for each cluster a set of features (conditions) that manifest the cluster. In this paper, we propose an ensemble clustering approach with feature selection and grouping for clustering high-dimensional biological big data. It uses two well-approved clustering methods: (a) k-means clustering and (b) similarity-based clustering. This approach selects the most relevant features in the dataset and grouping them into subset of features to overcome the problems associated with the traditional clustering methods. Also, we applied biclustering on each cluster that generated by ensemble clustering to find the sub-matrices in the biological data by the mean squared residue scores. We have applied the proposed clustering method on unlabeled genomic data (148 Exome datasets) of Brugada syndrome to discover previously unknown data patterns. Experiments verify that the proposed clustering method achieved high performance clustering results on high-dimensional biological big data.

Keywords: biclustering, biological big data, brugada syndrome, clustering, high-dimensional data.

1 INTRODUCTION

Biological big data mining is a challenging task to discover hidden patterns/knowledge in the data and handle the complexity of information with a reasonable accuracy. In general, the biological data is big (Petabyte even Exabyte), which represent the information of biological systems, including clinical and genomic data, molecular imaging and pathways, and different populations of people [1]. Biological big data can be especially useful in health care for disclosing genetic contributions to disease [2]. Biological data is much more heterogeneous and take less than a year to double in size [3]. The amount of genetic sequencing data from the Human Genome Project (HGP) turns into big data in every passing year [4]. Biological big data mining is a multidimensional view that follows: accuracy, completeness, consistency, timeliness, believability, and interpretability. Recently, data mining tools and techniques bring computational intelligent researchers into data analysis methods for analyzing biological big data that enable us to understand the basic biological/biomedical mechanisms and how the results can be applied in the future research of the bioinformatics field [5].

Clustering of high-dimensional big data is a difficult task. The most conventional clustering algorithms fail to generate meaningful results because of the inherent sparsity of the data space [6,7]. Many real-world data sets consist of a very high-dimensional feature space such as bioinformatics or web mining, where dimensionality of the feature space can be as high as a few thousands. For example, if a data set has 1,000 dimensions and we want to find clusters of dimensionality 10, then

This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-328-337

there are $\binom{1000}{10}$= 2.63×10^{23} possible subspaces [8]. Clustering of high-dimensional data can be classified into two categories: (a) subspace clustering, and (b) dimensionality reduction method. In high-dimensional data, subspace clustering tries to find clusters in existing subspaces using a subset of features in the full space [9], while the dimensionality reduction method constructs a new space instead of using subspaces of the original high-dimensional data. It tries to construct a much lower-dimensional space and search for clusters in such a space. It is effective in image processing, but computationally costly for big data.

In this paper, we have proposed an ensemble clustering with feature selection and grouping for clustering high-dimensional biological big data. The ensemble method uses k-means and similarity-based clustering. We have applied data pre-processing and feature selection techniques to select most relevant features in the data and grouping them into subset of features. We used unsupervised feature selection approach based on measuring similarity between features by maximum information compression index. Then, we cluster the biological data using ensemble clustering. Finally, the biclustering method is applied on each cluster that generated by ensemble clustering to find the sub-matrices in biological data by the mean squared residue scores. We have tested the proposed method on the 148 Exome unlabeled data sets to cluster the DNA variants for Brugada syndrome (BrS). BrS is a genetic disease that increases the risk of sudden cardiac death (SCD) at a young age.

The remaining of the paper is organized as follows. In Section 2, we review the k-means and similarity-based clustering. In Section 3, we present the ensemble clustering with biclustering algorithm. In Section 4, we demonstrate the performance of ensemble clustering. Section 5, concludes the paper with some remarks.

2 CLUSTER ANALYSIS

Clustering is the process of grouping instances into clusters so that instances within a cluster have high similarity in comparison to one another, but are very dissimilar to instances in other clusters [10]. Similarities and dissimilarities of instances are based on the predefined features of the data. Let X be the unlabeled data set, that is,

$$X = \{x_1, x_2, ...x_N\} \tag{1}$$

The partition of X into k clusters, $C_1, \ldots, C_k$, so that the following conditions are met in eqns (2)–(4).

$$Ci \neq \varnothing,\ i = 1,...,k \tag{2}$$

$$\cup_{i=1}^{k} C_i = X \tag{3}$$

$$C_i \cap C_j = \varnothing, i \neq j, i, j = 1, ..., k \tag{4}$$

A similarity measure (SM), sim(x_i, x_l), defined between any two instances, xi , xl ∈ X, and an integer value k, the clustering problem is to define a mapping $f : X \rightarrow 1, \cdots, k$, where each instance, x_i is assigned to one cluster C_i, $1 \leq i \leq k$. Given a cluster, C_i, $\forall x_{il}, x_{im} \in C_i$, and $x_j \notin C_i$, $sim(x_{il}, x_{im}) > sim(x_{il}, x_j)$, which together satisfy the following requirements: (1) each cluster must contain at least one instance, and (2) each instance must belong to exactly one cluster [11]. A distance measure (DM), $dis(x_i, x_l)$, where $x_i, x_l \in X$, as opposed to similarity measure. Let's consider the well-known Euclidean distance between two instances in Euclidean space in eqn (5).

$$dis(x_i, x_l)\sqrt{\sum_{i=1}^{m}(x_i - x_l)^2} \tag{5}$$

where, $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ and $x_l = (x_{l1}, x_{l2}, \cdots, x_{lm})$ are two instances in Euclidean m-space.

2.1 K-Means Clustering

The k-Means clustering defines the centroid of a cluster, *Ci* as the mean value of the instances $\{x_{i1}, x_{i2}, \cdots, x_{iN}\} \in C_i$. It proceeds as follows. First, it randomly selects k instances, $\{x_{k1}, x_{k2}, \cdots, x_{kN}\} \in X$ each of which initially represents a cluster mean/center. For each of the remaining instances, $x_{\mathrm{i}} \in X$, x_i is assigned to C_i to which it is more similar, based on the eqn (5) between the instance and the cluster mean. It then iteratively improves the within-cluster variation [12,13]. For each cluster, C_i, it computes the new mean using the instances assigned to the cluster in the previous iteration. All the instances, $x_i \in X$ are then reassigned into clusters using the updated means as the new cluster centers. The iterations continue until the assignment is stable. The cluster mean of $C_i = \{x_{i1}, x_{i2}, \cdots, x_{iN}\}$ is defined in eqn (6).

$$Mean = C_i = \frac{\sum_{j=1}^{N}(x_{ij})}{N} \tag{6}$$

In k-Means, the initial cluster means are assigned randomly. It is not guaranteed to converge to the global optimum and often terminates at a local optimum [14,15]. Algorithm 1 outlines the k-Means clustering method.

2.2 Similarity-Based Clustering

A similarity-based clustering method (SCM) is an effective and robust clustering approach based on the similarity of instances [16,17]. The instances in SCM can self-organize local optimal cluster number and volumes without using cluster validity functions. Let's consider sim(x_i, x_l) as the similarity measure between instances x_i and the lth cluster center x_l. The goal is to find x_l to maximize the total similarity measure as shown in eqn (7).

Algorithm 1 k-Means Clustering

```
Input: $X = \{x_1, x_2 \cdots, x_N\}$ // A set of unlabeled instances.
$k$ // the number of clusters
Output: A set of $k$ clusters.
Method:
    arbitrarily choose k number of instances, $\{x_{k1}, x_{k2}, \cdots, x_{kN}\} \in X$ as the initial $k$
      clusters center;
    repeat
      (re)assign each $x_i \in X \rightarrow k$ to which the $x_i$ is the most similar based on the mean
        value of the $x_m \in k$;
      update the $k$ means, that is, calculate the mean value of the instances for each cluster;
    until no change
```

$$J_s(C) = \sum_{l=1}^{k}\sum_{i=1}^{N} f(sim(x_i, x_l)) \tag{7}$$

Where, $f(sim(x_i, x_l))$ is a reasonable similarity measure and $C = \{C1, \cdots, Ck\}$.

In general, SCM uses feature values to check the similarity between instances. However, any suitable DM can be used to check the similarity between the instances. Algorithm 2 outlines the SCM.

3 CLUSTERING BIOLOGICAL BIG DATA

In this paper, we have proposed an ensemble clustering with feature selection and grouping method for clustering high-dimensional biological big data. To cluster the biological big data, we follow the steps as shown in Fig. 1.

Algorithm 2 Similarity-based Clustering

```
Input: $X = \{x1, x2, \cdots, xN\}$ // A set of unlabeled instances.
Output: A set of clusters, $C = \{C1, C2, \cdots, Ck\}$.
Method:
$C = \emptyset$;
$k = 1$;
$Ck = \{x1\}$;
$C = C \cup Ck$;
for $i = 2$ to $N$ do
    for $l = 1$ to $k$ do
        find the $l$th cluster center $x_l \in C_l$ to maximize the similarity measure, $sim(x_i, x_l)$;
    end for
    if $sim(xi, xl) \geq threshold\_value$ then
        $Cl = Cl \cup xi$
    else
        $k = k + 1$;
        $Ck = \{xi\}$;
        $C = C \cup Ck$;
    end if
end for
```

3.1 Data pre-processing

It transforms raw data into an understandable format for further processing, which includes several techniques: (a) data cleaning, (b) data integration, (c) data transformation, (d) data reduction, and (e) data discretization. *Data cleaning* is the process of dealing with missing values. *Data integration* merges data from different multiple sources into a coherent data store like data warehouse or integrate metadata. *Data transformation* includes the followings: (a) normalization, (b) aggregation, (c) generalization, and (d) feature construction. *Data reduction* obtains a reduced representation of data set (eliminating redundant features/ instances). *Data discretization* involves the reduction of a number of values of a continuous feature by dividing the range of feature intervals.

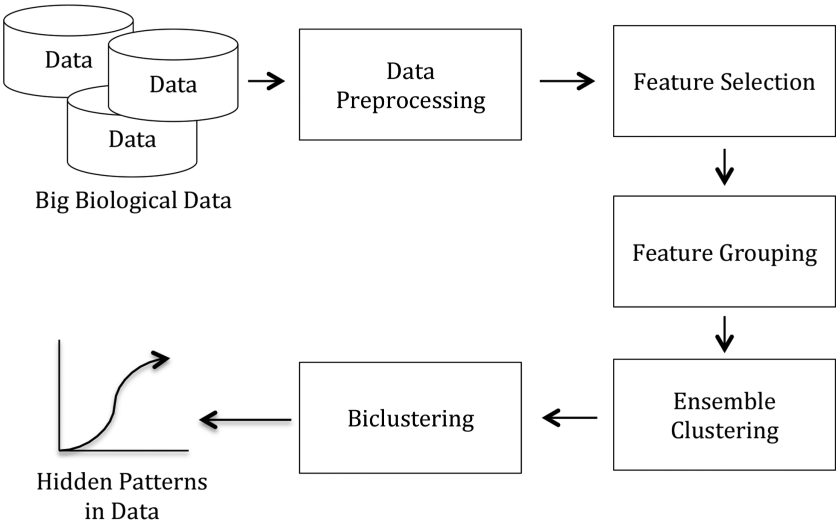

Figure 1: Pattern extracting process from biological big data.

3.2 Feature selection and grouping

Feature selection is the process of selecting a subset of relevant features *d* from a total of *D* original features for following three reasons: (a) simplification of models, (b) shorter training times, and (c) reducing overfitting [18]. It is a form of search based on a given optimization principal that improves the performance of the mining model. In biological data, features may contain false correlations and the information they add is contained in other features. In this paper, we have applied an unsupervised feature selection approach based on measuring similarities between features by maximum information compression index [19]. We have quantified the information loss in feature selection with entropy measure technique. After selecting the subset of features from the data, we have grouped them into two groups: nominal and numeric features.

3.3 Ensemble clustering

Ensemble clustering is a process of integrating multiple clustering methods to form a single strong clustering approach that usually provides better clustering results [20]. It performs more effectively in high-dimensional complex data. It generates a set of clusters from a given unlabeled data and then combines the clusters into final clusters to improve the quality of individual clustering. Generally, three strategies are applied in ensemble clustering: (a) using different clustering algorithms on the same data set to create heterogeneous clusters, (b) using different samples/subsets of the data with different clustering algorithms to cluster them to produce component clusters, and (c) running the same clustering algorithm many times on same data set with different parameters or initializations to create homogeneous clusters. The main goal of the ensemble clustering is to integrate component clustering into one final clustering with a higher accuracy. In this paper, we have used an ensemble clustering by employing k-means and similarity-based clustering. We have applied k-means algorithm on numeric features of data set and similarity-based algorithm on nominal features of data set. Finally, we have merged the clusters into final clusters.

3.4 Biclustering

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of instances, $A = \{a_1, a_2, \cdots, a_m\}$ be a set of features and $E = [e_{ij}]$ be a data matrix, where $1 \le i \le n$ and $1 \le j \le m$. A submatrix $I \times J$ is defined by a subset $I \subseteq X$ of instances and a subset $J \subseteq A$ of features. The mean of the *i*th row and the *j*th column for submatrix $I \times J$ is shown in eqns (8) and (9), respectively.

$$e_{iJ} = \frac{\sum_{j \in J} e_{ij}}{|J|} \tag{8}$$

$$e_{Ij} = \frac{\sum_{i \in I} e_{ij}}{|I|} \tag{9}$$

So, the mean of all elements in the submatrix $I \times J$ is shown in eqn (10).

$$e_{IJ} = \frac{\sum_{i \in I, j \in J} e_{ij}}{|I||J|} = \frac{\sum_{i \in I} e_{iJ}}{|I|} = \frac{\sum_{j \in J} e_{Ij}}{|J|} \tag{10}$$

We can define the *residue* score of *eij* in a submatrix *EI J* by eqn (11).

$$residue(e_{ij}) = e_{ij} - e_{iJ} - e_{Ij} + e_{IJ} \tag{11}$$

In a bicluster, the quality of a submatrix is measured by the *mean squared residue* score as shown in eqns (12) and (13).

$$H(I,J) = \frac{\sum_{i \in I, j \in J} (e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|I||J|} \tag{12}$$

$$H(I,J) = \frac{\sum_{i \in I, j \in J} (residue(e_{ij}))^2}{|I||J|} \tag{13}$$

If $H(I, J) \le \delta$, where $\delta \ge 0$ is a threshold, then the submatrix $I \times J$ is a δ-bicluster. The submatrix $I \times J$ will be a perfect bicluster with coherent values, if $\delta = 0$. We can specify the tolerance of average

Algorithm 3 δ-Biclustering

Input: E, a data matrix and $\delta \ge 0$, the maximum acceptable mean squared residue score.

Output: E_{IJ}, a δ-bicluster that is a submatrix of E with row set I and column set J, with a score no longer than δ.

Initialization: I and J are initialized to the instance and feature sets in the data and $EI\,J = E$.

Deletion phase:

```
1: compute eiJ for all $i \in I$, eIj for all $j \in J$, eIJ, and $H(I, J)$;
2: if $H(I, J) \le \delta$ then
3:     return EIJ;
4: end if
```

```
    find the rows $i \in I$ with $J$ $d(i) = \frac{\sum_{j \in J}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|J|}$;
    find the columns $j \in J$ with $d(i) = \frac{\sum_{i \in I}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|I|}$;
    remove rows $i \in I$ and columns $j \in J$ with larger $d$;
**Addition phase:**
    compute $eiJ$ for all $i$, $eIj$ for all $j$, $eIJ$, and $H(I, J)$;
    add the columns $j \notin J$ with $\frac{\sum_{i \in I}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|I|} \le H(I,J)$;
    recompute $eiJ$, $eIJ$ and $H(I, J)$;
    add the rows $i \notin I$ with $\frac{\sum_{j \in J}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|J|} \le H(I,J)$;
    **for** each row $i \notin I$ **do**
        **if** $\frac{\sum_{j \in J}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2}{|J|} \le H(I,J)$;
            add inverse of $i$;
        **end if**
    **end for**
    return $EIJ$;
```

noise per element against a perfect bicluster by setting $\delta > 0$. Algorithm 3 outlines the δ- biclustering method.

4 EXPERIMENTS

In this study, we have focused on BrS genetic disease for the molecular genetic clustering. We have used 148 unlabeled Exome data sets, which is the part of the genome formed by exons. An exon is any DNA sequence within a gene. Exome consists of all DNA that is transcribed into mature RNA. Each Exome data set contains 147 features (feature values are numeric and nominal). There are total 19,687 variants in 148 Exome data sets and among these variants 17,795 variants satisfy the gene panel of BrS. Finally, after data pre-processing, we have only 2,143 variants that satisfy BrS that is shown in Fig. 2. We have grouped the BrS variants into five clusters using the proposed ensemble clustering that are shown in Fig. 3. Then, we have applied δ-Biclustering using Algorithm 3 on each cluster to find the data pattern in Exome data sets.

5 CONCLUSIONS

No single clustering method is optimal. Different clustering methods may produce different clusters, because they impose different structure on data set. Ensemble clustering performs more effectively in high-dimensional biological data, and it is a good alternative when facing cluster analysis problems. In this paper, we have proposed an ensemble clustering using k-means and similarity-based

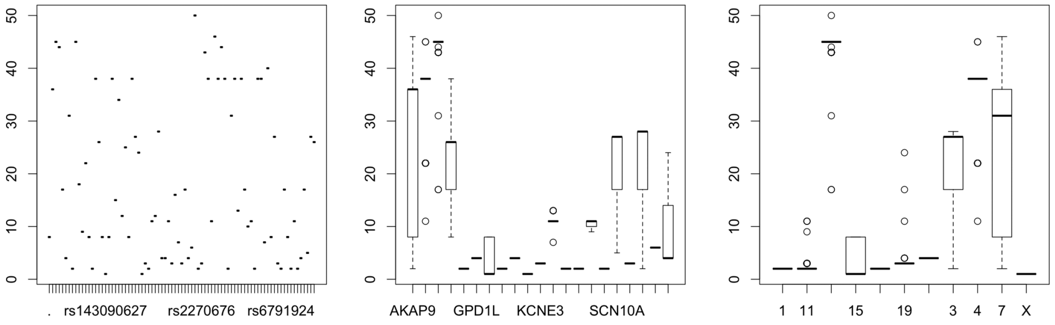

Figure 2: Unlabeled 148 Exome data sets of BrS.

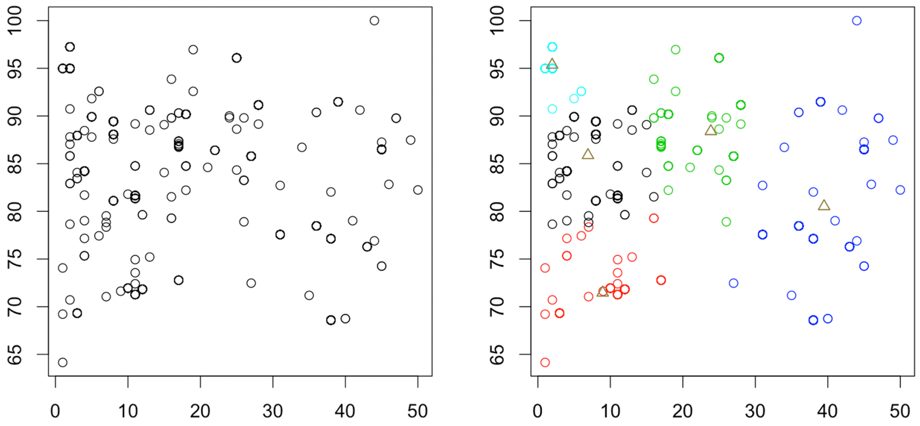

Figure 3: Distribution of BrS variants in clusters using proposed ensemble clustering.

clustering. Also, we have applied Biclustering on each cluster that is generated by ensemble clustering, not on the full data set. In future work, soft clustering will be associated with the ensemble model for clustering the biological data.

ACKNOWLEDGMENT

We appreciate the support for this research received from the BRiDGEIris (BRussels big Data platform for sharing and discovery in clinical GEnomics) project that is being hosted by IB^2 (Interuniversity Institute of Bioinformatics in Brussels) and funded by INNOVIRIS (Brussels Institute for Research and Innovation). Also, FWO research project G004414N “Machine Learning for Data Mining Applications in Cancer Genomics”.

REFERENCES

[1] Li, Y. & Chen, L., Big biological data: challenges and opportunities. *Genomics Proteomics Bioinformatics*, **12**(5), pp. 187–189, 2014. http://dx.doi.org/10.1016/j.gpb.2014.10.001

[2] May, M., Big biological impacts from big data. *Science*, **344**(6189), pp. 1298–1300, 2014. http://dx.doi.org/10.1126/science.344.6189.1298

[3] Marx, V., The big challenges of big data. *Nature*, **498**(7453), pp. 255–260, 2013.

http://dx.doi.org/10.1038/498255a

[4] Qin, Y., Yalamanchili, H.K., Qin, J., Yan, B. & Wang, J., The current status and challenges in computational analysis of genomic big data. *Big Data Research*, **2**(1), pp. 12–18, 2015. http://dx.doi.org/10.1016/j.bdr.2015.02.005

[5] Herland, M., Khoshgoftaar, T.M. & Wald, R., A review of data mining using big data in health informatics. *Journal of Big Data*, **1**(2), pp. 1–35, 2014. http://dx.doi.org/10.1186/2196-1115-1-2

[6] Jing, L., Ng, M.K. & Huang, J.Z., An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, **19**(8), pp. 1026–1041, 2007. http://dx.doi.org/10.1109/TKDE.2007.1048

[7] Jing, L., Tian, K. & Huang, J.Z., Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition*, **48**(11), pp. 3688–3702, 2015. http://dx.doi.org/10.1016/j.patcog.2015.05.006

[8] Han, J., Kamber, M. & Pei, J., *Data Mining Concepts and Techniques,* 3rd edn., Morgan Kaufmann, 2011.

[9] Zhu, L., Cao, L., Yang, J. & Lei, J., Evolving soft subspace clustering. *Applied Soft Computing*, **14**(B), pp. 210–228, 2014.

[10] Xu, R. & Wunsch, D., Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, **16**(3), pp. 645–678, 2005. http://dx.doi.org/10.1109/TNN.2005.845141

[11] Filippone, M., Camastra, F., Masulli, F. & Rovetta, S., A survey of kernel and spectral methods for clustering. *Pattern Recognition*, **41**(1), pp. 176–190, 2008. http://dx.doi.org/10.1016/j.patcog.2007.05.018

[12] Tsapanos, N., Tefas, A., Nikolaidis, N. & Pitas, I., A distributed framework for trimmed kernel k-means clustering. *Pattern Recognition*, **48**(8), pp. 2685–2698, 2015. http://dx.doi.org/10.1016/j.patcog.2015.02.020

[13] Malinen, M.I., Mariescu-Istodor, R. & Fr¨anti, P., K-means: clustering by gradual data transformation. *Pattern Recognition*, **47**(10), pp. 3376–3386, 2014. http://dx.doi.org/10.1016/j.patcog.2014.03.034

[14] Bagirov, A.M., Ugon, J. & Webb, D., Fast modified global k-means algorithm for incremental cluster construction. *Pattern Recognition*, **44**(4), pp. 866–876, 2011. http://dx.doi.org/10.1016/j.patcog.2010.10.018

[15] Tzortzis, G.F. & Likas, C.L., The global kernel k-means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks*, **20**(7), pp. 1181–1194, 2009. http://dx.doi.org/10.1109/TNN.2009.2019722

[16] Yang, M.S. & Wu, K.L., A similarity-based robust clustering method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(4), pp. 434–448, 2004. http://dx.doi.org/10.1109/TPAMI.2004.1265860

[17] Farid, D.M., Zhang, L., Hossain, A., Rahman, C.M., Strachan, R., Sexton, G. & Dahal, K., An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications*, **40**(15), pp. 5895–5906, 2013. http://dx.doi.org/10.4304/jait.4.3.129-135

[18] Farid, D.M. & Rahman, C.M., Mining complex data streams: discretization, attribute selection and classification. *Journal of Advances in Information Technology*, **4**(3), pp. 129–135, 2013. http://dx.doi.org/10.1016/j.eswa.2013.05.001

[19] Mitra, P., Murthy, C.A. & Pal, S.K., Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(3), pp. 301–312, 2002. http://dx.doi.org/10.1109/34.990133

[20] Iam-On, N., Boongoen, T., Garrett, S. & Price, C., A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(12), pp. 2396–2409, 2011. http://dx.doi.org/10.1109/TPAMI.2011.84