Australian Jour nal of Basic and Applied Sciences

2018 January; 12(1): pages 45-53
DOI: 10.22587/ajbas.2018.12.1.9

Research Article AENSI Publications

Cluster Ensemble Approach for High Dimensional Data

1M. Pavithra and 2Dr.R.M.S. Parvathi

1Assistant Professor, Department C.S.E, Jansons Institute of Technology, Coimbatore, India.
2Dean- PG Studies, Sri Ramakrishna Institute of Technology, Coimbatore, India.

Correspondence Author: M. Pavithra, Assistant Professor, Department C.S.E, Jansons Institute of Technology, Coimbatore, India.

Received date: 22 December 2017, **Accepted date:** 22 January 2018, **Online date:** 5 February 2018

Abstract

Data clustering is one of the essential tools for perceptive structure of a data set. It plays a crucial and initial role in machine learning, data mining and information retrieval. The intrinsic properties of the traditional algorithms intended for numerical data, can be employed to measure distance between feature vectors and cannot be directly applied for clustering of categorical data ,Wherever domain value are distinct haven't any ordering outlined. The final data partition generated by traditional algorithms, results in incomplete information and the core ensemble information matrix presents only cluster data point relations with many entries left unknown and disgrace the quality of the resulting cluster. This paper discusses one method of clustering a high dimensional dataset using dimensionality reduction and context dependency measures (CDM). First, the dataset is partitioned into a predefined number of clusters using CDM. Then, context dependency measures are combined with several dimensionality reduction techniques and for each choice the data set is clustered again. The results are combined by the cluster ensemble approach. Finally, the Rand index is used to compute the extent to which the clustering of the original dataset (by CDM alone) is preserved by the cluster ensemble approach. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature. Cluster ensembles can provide robust and stable solutions by leveraging the consensus across multiple clustering results, while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned. In this paper, we address the problem of combining multiple weighted clusters which belong to different subspaces of the input space. We leverage the diversity of the input clustering's in order to generate a consensus partition that is superior to the participating ones. Since we are dealing with weighted clusters, our consensus function makes use of the weight vectors associated with the clusters. The experimental results show that our ensemble technique is capable of producing a partition that is as good as or better than the best individual clustering. Experiments on three real data sets were conducted with three data generation methods and three consensus functions. The results have shown that the ensemble clustering with Fast Map projection outperformed the ensemble clustering with random sampling and random projection. The proposed approach has produced higher efficient clustering with negligible overlapping. We have used iris data set for the evaluation of the proposed approach and the results shows that the proposed method has produced efficient result than others. We proposed a soft feature selection procedure (called LAC) that assigns weights to features according to the local correlations of data along each dimension. Dimensions along which data are loosely correlated receive a small weight, which has the effect of elongating distances along that dimension.

Keywords: Data mining, Clustering, Cluster Ensemble, Semi Supervised Clustering, High Dimensional Data.

INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified (Sunita Jahirabadkar and Parag Kulkarni, 2013). Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses (Mehta, R.G., *et al*., 2014). Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity (Jain, A.K., *et al*., 2010). Clustering is a collection of data objects, similar to one another within the same cluster and are dissimilar to objects in the other clusters (Nenad Tomasev, 2013). Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown, and it is up to the clustering algorithm to discover acceptable classes. Similarly, the Bayesian approach also available for clustering data points where it uses certain rules and conditions to compute the distance between data points according to the probabilistic nature of data points

Citation: M. Pavithra and R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.

(Weiguo Liu, 2011). There are many other approaches available for clustering data points, but most of them suffer with the scalability in size and accuracy. Our requirement is to cluster data points which are in high dimension and more diverse in nature (Bini Tofflin. R1., 2014).

We have used real world data sets like Enron which is a data set which has information about YouTube and face book users (Strehl, A. and J. Ghosh, 2009). The real world data set can be used for clustering which has many number of dimensions, because a single user may have any number of contacts with various groups and each can be considered as a dimension. To identify or group similar interested users, the clustering approach can be used (Sunita Jahirabadkar and Parag Kulkarni, 2013). The Enron data set has number of users and each has connected with other users of the same group and from other groups. The entry in the data set has a pair of numbers which shows that there is a connection between two users of the network. If a user is interested in data mining then the second user is also has interested in data mining. This kind of data set is useful in grouping similar interested users of the network to share information between them and so on. The word high dimensional data means having number of attributes and also huge in size (Junming Shao, 2011).

To cluster such data we need to design a metric to compute distance between data points. There are many number of algorithms have been introduced for clustering categorical data, each has its own strengths and weaknesses (Mehta, R.G., *et al*., 2014). For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results (Weiguo Liu, 2011). The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering. For example the feature-based approach transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels), the direct approach that finds the final partition through relabeling the base clustering results, graph-based algorithms that employ a graph partitioning methodology and the pair wise-similarity approach that makes use of co-occurrence relations between data points (Bini Tofflin. R1., 2014). Cluster ensembles off er a solution to challenges inherent to clustering arising from its ill-posed nature. Cluster ensembles can provide more robust and stable solutions by leveraging the consensus across multiple clustering results, while averaging out emergent spurious structures that arise due to the various biases to which each participating algorithm is tuned (Jain, A.K., *et al*., 2010). In high dimensional data, clusters often exist in different subspaces. Ensemble clustering based on full space clustering algorithms fails to cluster such data. The innovation of subspace ensemble clustering techniques is promised to resolve this problem. Recently, two methods for generating low dimensional component data have been used to resolve the problem of subspace ensemble clustering of high dimensional data clustering (Sunita Jahirabadkar and Parag Kulkarni, 2013). One method generates low dimensional data by randomly sampling different features. The other method generates low dimensional component data by using a random projection matrix to project the original high dimensional data onto a low dimensional space. We call the former random sampling method and the latter random projection method. In recent days, different flavors of random projection are available (Song, Q., *et al*., 2011; Mehta, R.G., *et al*., 2014) but for ensemble clustering former random projection has been used (Junming Shao, 2011). Both, random sampling and random projection benefit ensemble clustering for high dimensional sparse data (Bini Tofflin. R1., 2014). However, the drawback of these methods is that they cannot well preserve the clustering structure of the original data in their generated low dimensional component data, which increases discrepancy of clustering structures in component data sets, thus affecting the performance of ensemble clustering for high dimensional data (Strehl, A. and J. Ghosh, 2009). Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering (Jain, A.K., *et al*., 2010). The second finding is that the ensemble process exposes weakness in preceding assumptions. The third finding is that the practice of establishing eligible voters and privileged voters improves the quality and the interpretation of clusters (Sunita Jahirabadkar and Parag Kulkarni, 2013). Thus, using cluster ensembles addresses the three main barriers to wide scale adoption. Aligned with these areas of concern are the three core findings and experimental results in support of them (Nenad Tomasev, 2013).

The proposed model combines the three techniques, subspace clustering, H-K clustering and ensemble clustering and their advantages to improve the performance of clustering result on high dimensional data which will simultaneously overcome the limitations of H-K clustering algorithm for high dimensional data (as high computational complexity and poor accuracy) (Junming Shao, 2011). Creating clustering algorithm that can effectively deal with high dimensional data is not an easy task (Weiguo Liu, 2011; Sunita Jahirabadkar and Parag Kulkarni, 2013). By decreasing the dimension of the data, some of the researchers have recently solved the high dimensional problem (Nenad Tomasev, 2013; Junming Shao, 2011). As, poor classification efficiency due to high dimension of the feature space is the bottleneck of the classification task, dimensionality reduction is of great significance for the quality and efficiency of a classifier (Bini Tofflin. R1., 2014), particularly for large scale real-time data. Conventional and modern dimensionality reduction techniques can be broadly classified into Feature Extraction (FE) (Jain, A.K., *et al*., 2010) and Feature Selection (FS) (Nenad Tomasev, 2013; Mehta, R.G., *et al*., 2014) methods. FE methods are normally more effective than the FS techniques (except for a few specific cases) and their high effectiveness for real-world dimensionality reduction applications has been verified already (Ying he, *et al*., 2013).

Related Work:

Clustering for High Dimensional Data, Density based Subspace Clustering Algorithms (Sunita Jahirabadkar and Parag Kulkarni, 2013), treat clusters as the dense regions compared to noise or border regions. Many momentous density based subspace clustering algorithms exist in the literature (Nenad Tomasev, 2013). Each of them is characterized by different characteristics caused by different assumptions, input parameters or by the use of different techniques etc. Hence it is quite unfeasible for the future developers to compare all these algorithms using one common scale (Song, Q., *et al*., 2011). This paper presents a review of various density based subspace clustering algorithms together with a comparative chart focusing on their distinguishing characteristics such as overlapping / non-overlapping, axis parallel / arbitrarily oriented and so on (Junming Shao, 2011).

The Role of Hubness in Clustering High Dimensional Data (Nenad Tomasev, 2013), show that hubness, i.e., the tendency of high-dimensional data to contain points (hubs) that frequently occur in k-nearest neighbour lists of other points, can be successfully exploited in clustering. The algorithm validated the hypothesis by demonstrating that hubness is a good measure of point centrality within a high dimensional data cluster (Bini Tofflin. R1., 2014). It also proposes several hubness-based clustering algorithms, showing that major hubs can be used effectively as cluster prototypes or as guides during the search for centroid-based cluster configurations (Jain, A.K., *et al*., 2010).

Detection of Arbitrarily Oriented Synchronized Clusters in High-Dimensional Data (Junming Shao, 2011), propose ORSC (Arbitrarily Oriented Synchronized Clusters), a novel effective and efficient method to subspace clustering inspired by synchronization is a basic phenomenon prevalent in nature, capable of controlling even highly complex processes such as opinion formation in a group (Mehta, R.G., *et al*., 2014). Control of complex processes is achieved by simple operations based on interactions between objects. Relying on the interaction model for synchronization, the proposed approach ORSC naturally detects correlation clusters in arbitrarily oriented subspaces, including arbitrarily shaped non-linear correlation clusters. ORSC approach is robust against noise points and outliers (Strehl, A. and J. Ghosh, 2009).

Clustering Algorithm for High Dimensional Data Stream over Sliding Windows (Weiguo Liu, 2011), proposes an effective clustering algorithm referred as HSWStream for high dimensional data stream over sliding windows (Bini Tofflin. R1., 2014). This algorithm handles the high dimensional problem with projected clustering technique. It deals with the in cluster evolution with exponential histogram of cluster feature called EHCF and eliminates the influence of old points with the fading temporal cluster features. A fast computational method is employed to maintain the exponential histogram cluster feature. The algorithm produces high quality clusters with less memory usage (Nenad Tomasev, 2013).

PROCLUS (Mehta, R.G., *et al*., 2014), is an efficient high dimensional clustering algorithm; consist of significant issues like inconsistency in results and expert supervised subspaces. MPROCLUS: a modified PROCLUS algorithm is proposed, aimed. A Relevant Clustering Algorithm for High- Dimensional Data (Bini Tofflin. R1., 2014), is used for finding the subset of features. A Relevant clustering algorithm renders efficiency and effectiveness to find the subset of features (Sunita Jahirabadkar and Parag Kulkarni, 2013). Relevant clustering algorithm work can be done in three steps. First step involves elimination of irrelevant features from the dataset; the relevant features are selected by the features having the value greater than the predefined threshold. In the second step, selected relevant features are used to generate the graph, divide the features using graph theoretic method, and then clusters are formed by using Minimum Spanning Tree. In the third step, the subsets features that are more related to the target class are selected (Ying he, *et al*., 2013).

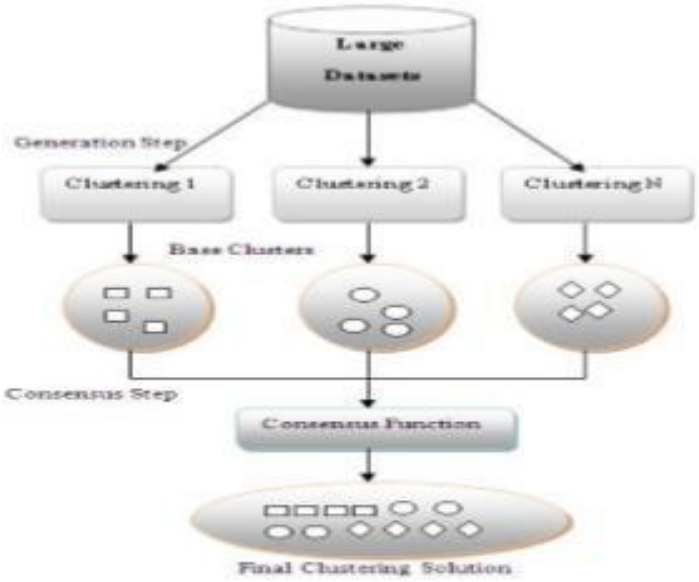
Global dimensionality reduction techniques are unable to capture local correlations of data. Thus, a proper feature selection procedure should operate locally in input space. Local feature selection allows one to embed diff erent distance measures in diff erent regions of the input space; such distance metrics reflect local

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.

correlations of data. In (Weiguo Liu, 2011) we proposed a soft feature selection procedure (called LAC) that assigns weights to features according to the local correlations of data along each dimension. Dimensions along which data are loosely correlated receive a small weight, which has the eff ect of elongating distances along that dimension (Tidke, B.A., *et al*., 2012). Features along which data are strongly correlated receive a large weight, which has the eff ect of constricting distances along that dimension. Thus the learned weights perform a directional local reshaping of distances which allows a better separation of clusters, and therefore the discovery of diff erent patterns in diff erent subspaces of the original input space (Reza Ghaemi, 2009).

Cluster Ensemble Methodology:

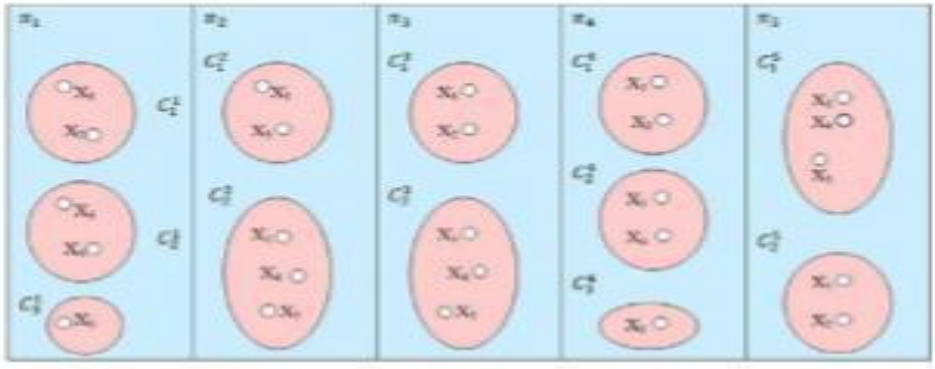
Clustering Ensemble was mainly proposed to overcome the lack of cluster quality resulted from the individual clustering (Jain, A.K., *et al*., 2010) algorithms. This eminence leads to the emergence of several cluster ensemble techniques over the past decades. The cluster ensemble paradigm comprises of two main aspects, first phase is to produce the several clustering membership and the second phase is to merge the clusters into a global design of ultimate partition. The general process of the cluster ensemble methodology (Sunita Jahirabadkar and Parag Kulkarni, 2013) is shown.



The method of clustering ensemble process initiates by generating diverse population of the clustering partitions through several generative mechanisms. Ensembles are more efficient, when assembled from a set of forecaster whose errors are dissimilar (Song, Q., *et al*., 2011). To a massive extent, diversity among the ensemble methods will enhances the result of cluster ensemble. In particular the results obtained from clustering the dataset using any single clustering algorithm over much iteration are usually similar to each other (Tidke, B.A., *et al*., 2012).

Creating Cluster Ensemble:

Consider the Dataset be a set of data points and π denotes the cluster ensembles such that are the ensemble members with base clustering. Each base clustering profits a set of clusters whereas is number of clusters in the clustering results (Reza Ghaemi, 2009).



Proposed Work:

V a. Locally Adaptive Clustering:

Let us consider a set of n points in some space of dimensionality D.A weighted cluster C is a subset of data points, together with a vector of weights w = (w1,...,wD)t, such that the points in C are closely clustered according to the L2 norm distance weighted using w (Zhao Yanchang *et al*., 2009). The component wj measures the degree of correlation of points in C along feature j. The problem is how to estimate the weight vector w for each cluster in the dataset. In traditional clustering, the partition of a set of points is induced by a set of representative vectors, also called centroids or centres. The partition induced by discovering weighted clusters (Ying he, *et al*., 2013).

$$S_j = \{\mathbf{x} | (\sum_{i=1}^{D} w_{ji}(x_i - c_{ji})^2)^{1/2} < (\sum_{i=1}^{D} w_{li}(x_i - c_{li})^2)^{1/2}, \forall l \neq j\}, j = 1, \ldots, k$$

We point out that LAC has shown a highly competitive performance with respect to other state-of-the-art subspace clustering algorithms (Weiguo Liu, 2011). Therefore, improving upon LAC performance is a desirable achievement.

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.

$$w_{ji}^{*} = \frac{\exp(-X_{ji}/h)}{\sum_{i=1}^{D} \exp(-X_{ji}/h)}$$

$$c_{ji}^{*} = \frac{1}{|S_j|} \sum_{\mathbf{x} \in S_j} x_i$$

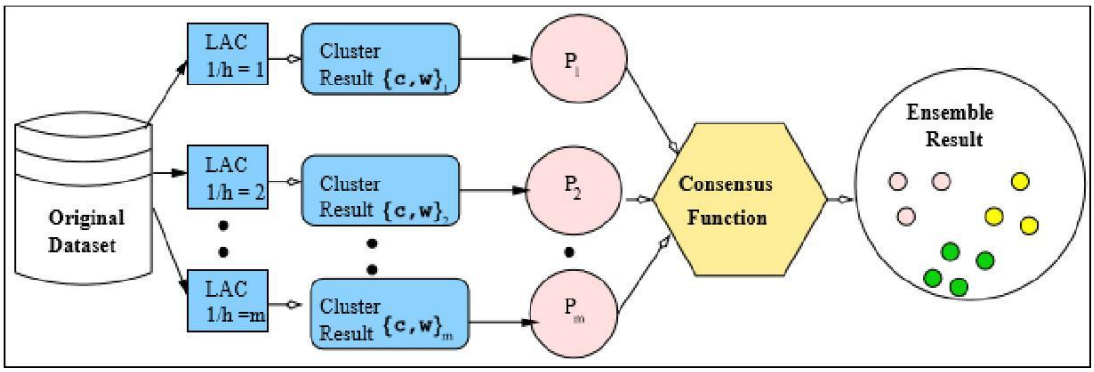


Fig. 1: The clustering ensemble process

V b. Weighted Similarity Partitioning Algorithm (Wspa):

LAC outputs a partition of the data, identified by the two sets {c1,...,ck} and {w1,...,wk}. Our aim here is to generate robust and stable solutions via a consensus clustering method (Sunita Jahirabadkar and Parag Kulkarni, 2013). We can generate contributing clustering's by changing the parameter h (as illustrated in Figure 1). The objective is then to find a consensus partition from the output partitions of the contributing clustering's, so that an "improved" overall clustering of the data is obtained (Weiguo Liu, 2011).

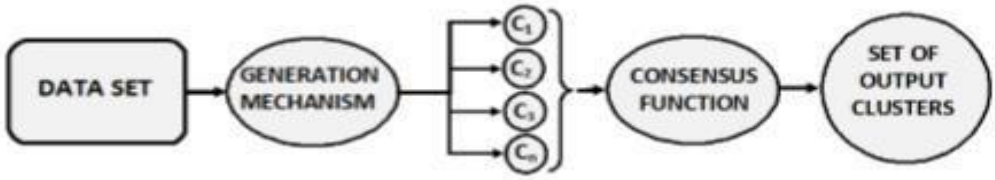
$$d_{il} = \sqrt{\sum_{s=1}^{D} w_{ls}(x_{is} - c_{ls})^2}$$

Since LAC produces weighted clusters, we need to design a consensus function that makes use of the weight vectors associated with the clusters (Bini Tofflin. R1., 2014). The details of our approach are as follows. For each data point xi, the weighted distance from cluster Cl is given by

$$P(C_l|\mathbf{x}_i) = \frac{D_i - d_{il} + 1}{kD_i + k - \sum_l d_{il}}$$

$$P_i = (P(C_1|\mathbf{x}_i), P(C_2|\mathbf{x}_i), \ldots, P(C_k|\mathbf{x}_i))^t$$

We provide a nonparametric estimation of such probabilities based on the data and on the clustering result. We do not make any assumption about the specific form (e.g., Gaussian) of the underlying data distributions, thereby avoiding parameter estimations of models, which are problematic in high dimensions when the available data are limited (Jain, A.K., *et al*., 2010).



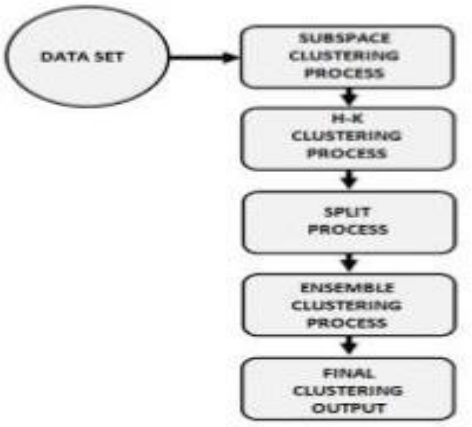
V c. H-K Clustering:

H-K clustering algorithm is proposed and implemented for deciding the k clusters for k-means algorithm (Weiguo Liu, 2011). It is implemented in divisive H-K and agglomerative H-K clustering. Divisive H-K algorithm implements a top-down approach which splits the whole dataset into the small clusters (Jain, A.K., *et al*., 2010). It divides the K clusters into K+1 clusters using K-means method. Agglomerative clustering works by merging the small clusters together. It merges the K clusters into K-I cluster (Jain, A.K., *et al*., 2010).

V c1. The H-K Clustering Process:

Adopt the H-K clustering on the k subspaces which is the output of stage 2. Apply the divisive HK means clustering, on the k subspaces it divides the k cluster dataset into k+1 clusters using k means method (Tidke, B.A., *et al*., 2011). This will help pick up the two elements that are furthest from each other in this cluster, so as to divide the distance between the two into 3 equivalent parts to produce one more new cluster (Zhao Yanchang *et al*., 2009). Repeat this process for the range of [2, k+10] and by applying the random selection method select the L cluster as the output and store them as H(1),H(2),H(3). . . . . , H(L).

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.



V d. Randomized Embedding Cluster Ensembles (RE-Clust):

Consider a data set X ={x1,x2,...,xn}, where xi ∈Rd, (1≤ i ≤ n); a subset A ⊆ {1,2,...,n} univocally individuates a subset of examples {xjj ∈ A} ⊆ X. The data set X may be represented as a d×n matrix D, where columns correspond to the examples, and rows correspond to the "components" of the examples x ∈ X. A k-clustering C of X is a list C =< A1,A2,...,Ak >, with Ai ⊆ {1,2,...,n} and such that SAi = {1,...,n}. A clustering algorithm C is a procedure that, having as input a data set X and an integer k, outputs a k-clustering C of X: C(X,k) =< A1,A2,...,Ak > (Weiguo Liu, 2011). The main ideas behind the proposed cluster ensemble algorithm RE-Clust (acronym for Randomized Embedding Clustering) are based on data compression, and generation and combination of multiple" base" clustering's (Bini Tofflin. R1., 2014). Indeed at first data are randomly projected from the original to lower dimensional subspaces, using projections described in Sect 2.2 in order to approximately preserve the distances between the examples. Then multiple clustering's performed on multiple instances of the projected data, and a similarity matrix between pairs of examples is used to combine the multiple clustering's (Jain, A.K., *et al*., 2010).

V d1. Re-Clust

Algorithm: Input:

– A data set X ={x1, x2... xn}, represented by a d×n D matrix.
– An integer k (number of clusters)
– A real ² > 0 (distortion level)
– An integer c (number of clusterings)
– Two clustering algorithms C and Ccom
– A procedure that realizes a randomized map µ

Begin algorithm:

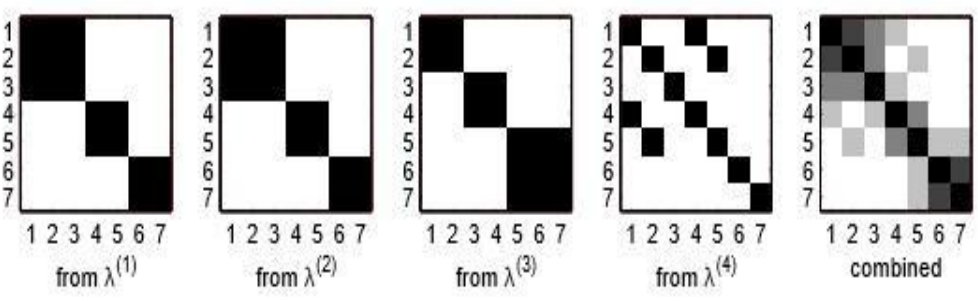
(1) D0 = 2·³2logn+logc ²2 ´
(2) For each i,j ∈{1,...,n} do Mij = 0
(3) Repeat for t = 1 to c
(4) Pt = Generate projection matrix (d, d0)
(5) Dt = Pt ·D
(6) < C(t) 1 ,C(t) 2 ,...,C(t) k >=C(Dt,k)
(7) For each i,j ∈{1,...,n} M(t) ij = 1 kPk s=1 I(i∈C(t) s)·I(j ∈C(t) s) end repeat
(8)M = Pc t=1 M (t) c
(9) < A1, A2... Ak >=Ccom (M, k)
End algorithm.

Output:

– The final clustering C =< A1, A2... Ak >

V e. Cluster-Based Similarity Partitioning Algorithm (CSPA):

Essentially, if two objects are in the same cluster then they are considered to be fully similar, and if not they are dissimilar. This is the simplest heuristic and is used in the Cluster-based Similarity Partitioning Algorithm (CSPA). With this viewpoint, one can simply reverse engineer a single clustering into a binary similarity matrix. Similarity between two objects is 1 if they are in the same cluster and 0 otherwise. For each clustering, a n×n binary similarity matrix is created. The entry-wise average of r such matrix representing their sets of groupings yields an overall similarity matrix.



VI. Experiments:

In this section, we empirically demonstrate that our proposed semi-supervised clustering algorithm is both efficient and eff ective.

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.

VI a. Datasets:

The data sets used in our experiments include six UCI data sets1. Here is some basic information of those data sets. Table 5 summarizes the basic information of those data sets.

• Balance. This data set was generated to model psychological experimental results. There are totally 625 examples that can be classified as having the balance scale tip to the right, tip to the left, or be balanced.

• Iris. This data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

• Ionosphere. It is a collection of the radar signals belonging to two classes. The data set contains 351 objects in total, which are all 34-dimensional.

• Soybean. It is collected from the Michalski's famous soybean disease databases, which contains 562 instances from 19 classes.

• Wine. The purpose of this data set is to use chemical analysis for determining the origin of wines. It contains 178 instances from 3 classes.

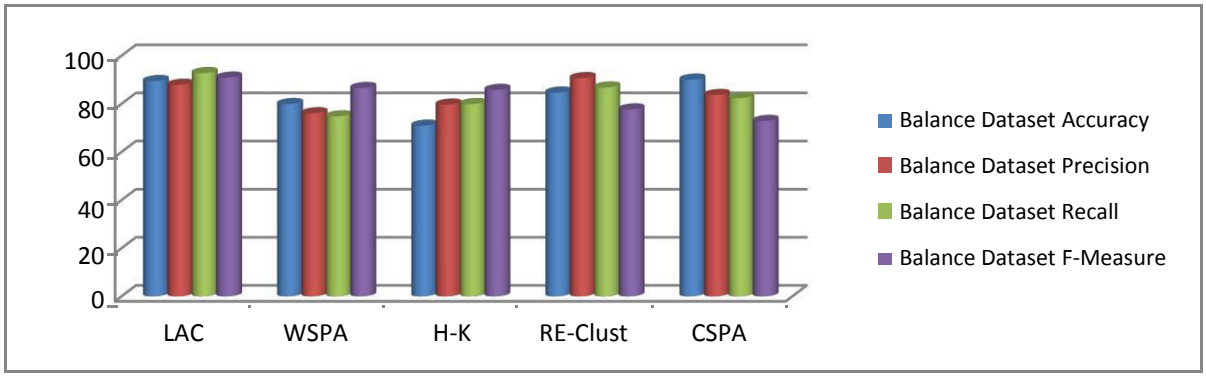
• Sonar. This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network, which contains 208 instances from 2 classes.

Datasets	Size	Classes	Dimensions
Balance	625	3	4
Iris	150	3	4
Ionosphere	351	2	34
Soybean	562	19	35

VII. Experimental Results:

VII a. Balance Dataset Results:

	Balance Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
LAC	89.45	87.91	92.77	90.89
WSPA	79.91	76.08	74.78	86.56
H-K	70.92	79.67	79.89	85.78
RE-Clust	84.67	90.67	86.78	77.67
CSPA	90.07	83.66	82.33	72.88

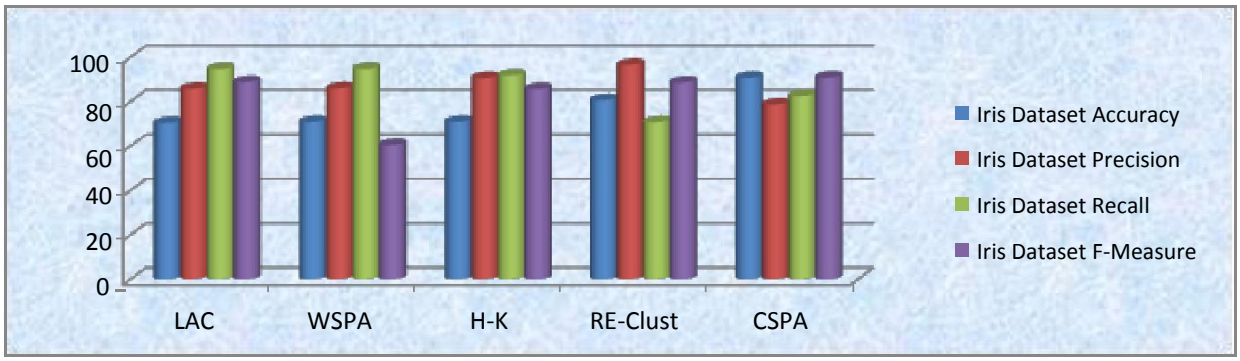


The above graph shows that performance of Balance dataset. The Accuracy of CSPA algorithm is 90.07 which is higher when compare to other four (LAC, WSPA, H-K, RE-Clust) algorithms. The Precision of RE-Clust algorithm is 90.67 which is higher when compare to other four (LAC, WSPA, H-K, CSPA) algorithms. The Recall of LAC algorithm is 92.77 which is higher when compare to other four (CSPA, WSPA, H-K, RE-Clust) algorithms. The F-Measure of LAC algorithm is 90.89 which is higher when compare to other four (CSPA, WSPA, H-K, RE-Clust) algorithms.

VII b. Iris Dataset Results:

	Iris Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
LAC	70.45	85.91	94.77	88.89
WSPA	70.91	86.08	94.78	60.56
H-K	70.92	90.67	91.89	85.78
RE-Clust	80.67	96.67	70.78	88.67
CSPA	90.78	78.76	82.54	90.89

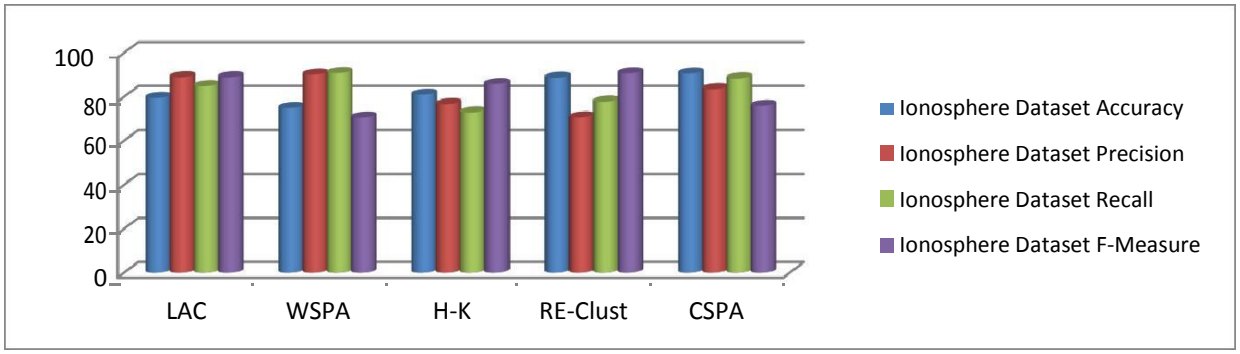
Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.



The above graph shows that performance of Iris dataset. The Accuracy of CSPA algorithm is 90.78 which is higher when compare to other four (LAC, WSPA, H-K, RE-Clust) algorithms. The Precision of RE-Clust algorithm is 96.67 which is higher when compare to other four (LAC, WSPA, H-K, CSPA) algorithms. The Recall of WEPA algorithm is 94.78 which is higher when compare to other four (LAC, CSPA, H-K, RE-Clust) algorithms. The F-Measure of CSPA algorithm is 90.89 which is higher when compare to other four (LAC, WSPA, H-K, RE-Clust) algorithms.

VII c. Ionosphere Dataset Results:

	Ionosphere Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
LAC	79.45	88.91	84.77	88.89
WSPA	74.91	90.08	90.78	70.56
H-K	80.98	76.67	72.89	85.78
RE-Clust	88.67	70.67	77.78	90.67
CSPA	90.56	83.45	88.34	75.89



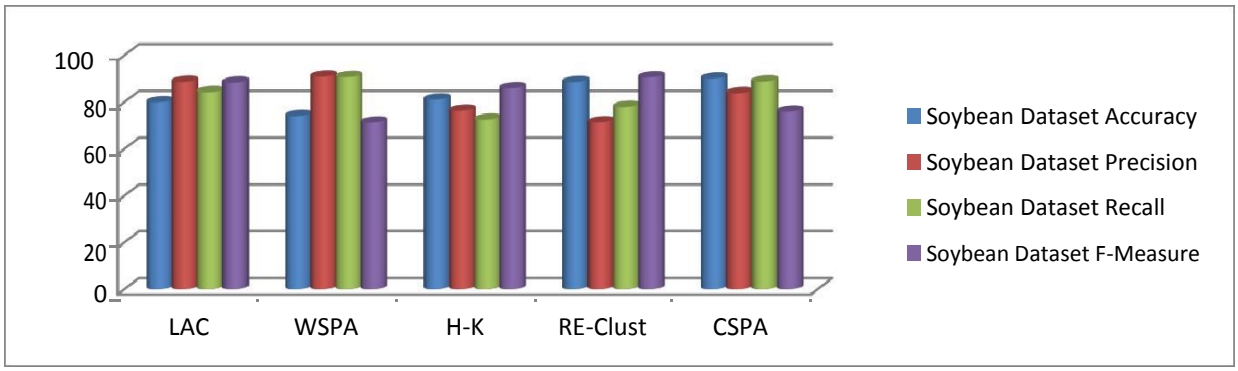
The above graph shows that performance of Ionosphere dataset. The Accuracy of CSPA algorithm is 90.56 which is higher when compare to other four (LAC, WSPA, H-K, RE-Clust) algorithms. The Precision of WSPA algorithm is 90.08 which is higher when compare to other four (LAC, CSPA, H-K, RE-Clust) algorithms. The Recall of WSPA algorithm is 90.78 which is higher when compare to other four (LAC, CSPA, H-K, RE-Clust) algorithms. The F-Measure of RE-Clust algorithm is 90.67 which is higher when compare to other four (LAC, WSPA, H-K, CSPA) algorithms.

VII d. Soybean Dataset Results:

	Soybean Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
LAC	79.89	88.65	84.23	88.34
WSPA	74.03	90.89	90.67	71.23
H-K	81.08	76.32	72.45	85.9
RE-Clust	88.54	71.32	77.89	90.56
CSPA	90.08	83.78	88.78	75.9

The above graph shows that performance of Soybean dataset. The Accuracy of CSPA algorithm is 90.08 which is higher when compare to other four (LAC, WSPA, H-K, RE-Clust) algorithms. The Precision of WSPA algorithm is 90.89 which is higher when compare to other four (LAC, CSPA, H-K, RE-Clust) algorithms. The Recall of WSPA algorithm is 90.67 which is higher when compare to other four (LAC, CSPA, H-K, RE-Clust) algorithms. The F-Measure of RE-Clust algorithm is 90.56 which is higher when compare to other four (LAC, WSPA, H-K, CSPA) algorithms.

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.



Conclusion:

Cluster ensembles have emerged as an efficient answer that's able to overcome the limitation of grouping mixed information, and develop the strength yet because the quality of cluster results. The most objective of cluster ensembles are to affix completely different cluster selections in such some way on accomplish accuracy bigger to it of any person cluster (Weiguo Liu, 2011). Cluster ensemble approach to categorical information cluster, transforms the first categorical information matrix to associate information-preserving graph partitioning technique may be directly applied to induce the ultimate information partition (Sunita Jahirabadkar and Parag Kulkarni, 2013). We have introduced two cluster ensemble techniques for the LAC algorithm. The experimental results show that our weighted clustering ensembles can provide solutions that are as good as or better than the best individual clustering, provided that the input clustering's are diverse. The proposed model provides a solution algorithm for processing the high dimensional dataset which is a combination of the three approaches and makes use of the advantages of ensemble and subspace clustering and simultaneously overcomes the limitations of the traditional H-K clustering such as, high computational complexity and poor accuracy by providing a three stage clustering process (Mehta, R.G., *et al*., 2014). Each algorithm has its own strengths and weaknesses (Strehl, A. and J. Ghosh, 2009). A number of subsequent methods have also been designed for determining multiple relevant subspaces for a candidate outlier, and then combining the results from different subspaces in order to create a more robust ensemble-based ranking. It is also possible to determine the outliers in arbitrarily oriented subspaces of the data (Song, Q., *et al*., 2011). Such methods are able to exploit the local correlations in the data in order to determine relevant outliers. Outlier analysis is the most difficult problem among all classes of subspace analysis problems. This led us to develop cluster ensembles, an approach to adopt multi-learner systems for clustering (Nenad Tomasev, 2013). We proposed a formal cluster ensemble problem and developed three effective and efficient algorithms to solve it. This combiner framework is useful in a variety of applications besides knowledge reuse (Junming Shao, 2011). Then using evolutionary process, it generates a better clustering solution by select a diverse set of clustering solution and then selecting high-quality clusters to derive a final clustering solution. The ensemble members can be built so as to take into account possible large dimensionality of the feature space. The solution proposed in (Jain, A.K., *et al*., 2010) is to use random linear projections in lower-dimensional spaces and run the clustering in these spaces. Unlike data size scalability, feature size scalability can be incorporated in the proposed variant (Bini Tofflin. R1., 2014). Combining advantages of RE-CLUST and divide and conquer strategy can help us in both efficiency and quality. Besides simulating of H-K is possible with recursive intrinsic divide and conquer method and creating nested clusters. HC algorithm can construct structured clusters (Junming Shao, 2011). Although H-K yields high quality clusters but its complexity is quadratic and is not suitable for huge datasets and high dimension data. In contrast RE-CLUST is linear with size of data set and dimension and can be used for big datasets that yields low quality. In this paper we present a method to use both advantages of H-K and RE-CLUST by introducing equivalency and compatible relation concepts. By these two concepts we defined similarity and our space and could divide our space by a specific criterion (Nenad Tomasev, 2013). Many directions exist to improve and extend the proposed method. Different applications can be used and examined the framework. Data mining is an interesting arena. Based on this method data stream processing can be improved. Data type is another direction to examine this method. In this study RE-CLUST has been used for second phase whereas we can use other clustering algorithms e.g. genetic algorithm, H-K algorithm, Ant clustering (Strehl, A. and J. Ghosh, 2009), Self Organizing Maps (Jain, A.K., *et al*., 2010), etc. Determining number of sub spaces can be studied as important direction for the proposed method.

Future Work:

In our future work we will consider utilizing our consensus function as a similarity matrix for hierarchical and spectral clustering. This approach will eliminate the requirement for balanced clusters (Song, Q., *et al*., 20111). We will extend our approach to be used with any subspace clustering technique. In addition, we aim at designing an ensemble that preserves a subspace clustering structure (Sunita Jahirabadkar and Parag Kulkarni, 2013). One possibility is to leverage the weight vectors associated with the input clustering that shares the highest NMI with the clustering produced by the ensemble (this can be performed using the RAND statistic) (Junming Shao, 2011). Another possibility is to infer a set of dimensions for each cluster from the clustering result of the ensemble.For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. (Nenad Tomasev, 2013) Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Cluster ensemble has proved to be a decent alternative when facing cluster analysis problems. To overcome such type of a problem we use the cluster ensemble method. It proved solution when facing such problems. Cluster ensemble is a process combining base clustering result into one final cluster. The aim of this method to improve robustness and quality of data clustering (Jain, A.K., *et al*., 2010). It is expected that the accuracy of the ensemble improves when a larger number of input clustering's is given, provided that the contributing 'clustering's diverse (Weiguo Liu, 2011). There are several directions for future research on relationship-based learning frameworks. This section highlights some of the promising directions. We will first discuss some theoretical and algorithmic improvements of cluster ensembles and conclude with a selection of particularly interesting application domains (Mehta, R.G., *et al*., 2014). These promising initial results invite further work, in particular in the application of other dimensionality reduction schemes and more complex ensemble combination rules, as well as in understanding how ensembles can be used for mitigating the tradeoff between denoising and feature preservation properties. The application of the proposed approach to larger scale data sets can also be the subject of future work, together with experimental evaluation (Song, Q., *et al*., 2011). The representations are fed into the initial clustering algorithm, and the effect of the initial clustering gives multiple partitions of the each representation. For each partition, we have applied various clustering validation standards for evaluating each partition received from the initial clustering and then we calculate the final weight for finding the final clustering (Weiguo Liu, 2011). Proposed approach can be highly effective to generate an initial clustering result with an automatically detected number of clusters, there are still many obvious directions to be explored in the future. Complexity of merging algorithm is high and needs to be making more efficient (Mehta, R.G., *et al*., 2014). In future work, other clustering algorithms for large scale dataset with mixed attribute types can be explored, also some weighting schemes on existing algorithms to perform well on their corresponding type of attributes to improve the proposed framework [10]. Clustering results on the categorical and numeric dataset are combined as a categorical dataset, which allows integration of other algorithms to produce corresponding clusters which leads to a better clustering accuracy (Tidke, B.A., *et al*., 2012).

REFERENCES

Song, Q., J. Ni, G. Wang, 2011. "A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions on Knowledge and Data Engineering.

Sunita Jahirabadkar and Parag Kulkarni, 2013." Article: Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms", International Journal of Computer Applications, 63(20) 29-35, February.

Citation: M. Pavithra and Dr.R.M.S. Parvathi, 2018. Cluster Ensemble Approach for High Dimensional Data.
Australian Journal of Basic and Applied Sciences., 12(1): 45-53.

Nenad Tomasev, 2013. "The Role of Hubness in Clustering High-Dimensional Data ", IEEE transactions on knowledge and data engineering, revised January.

Junming Shao, 2011." Detection of Arbitrarily Oriented Synchronized Clusters in High-Dimensional Data", International conference on data mining, pp: 607-616.

Weiguo Liu, 2011. "Clustering Algorithm for High Dimensional Data Stream over Sliding Windows, Trust", Security and Privacy in Computing and Communications (TrustCom), pp: 1537-1542.

Mehta, R.G., N.J. Mistry and M. Raghuwanshi, 2014."Article: Towards Unsupervised and Consistent High Dimensional Data Clustering", International Journal of Computer Applications, 87(2): 40-44.

Bini Tofflin. R1., 2014." A Relevant Clustering Algorithm for High- Dimensional Data ", International Journal of Innovative Research in Computer and Communication Engineering, 2, Special Issue 1.

Strehl, A. and J. Ghosh, 2009."Cluster ensembles — a knowledge reuse framework for combining multiple partitions", J. Mach. Learn. Res., 3: 583-617.

Jain, A.K., M.N. Murty P.J. Flynn, 2010. "Data Clustering: A Review," ACM Computing Surveys, 31(3): 264-323.

Ying he, Jian wang, Liang-xi, Lin Mei, Yan-feng Shang, Wen-fei Wang, 2013. "An h-k clustering algorithm based on ensemble learning", ICSSC.

Tidke, B.A., R.G. Mehta, D.P Rana, 2012. "A novel approach for high dimensional data clustering", ISSN: 2250–3676, [IJESAT] international journal of engineering science & advanced technology, 2(3): 645-651.

Reza Ghaemi, 2009."A Survey: Clustering Ensembles Techniques", proceedings of world academy of science, engineering and technology, 38, ISSN: 2070-3740.

Yanchang, Z., & Junde, S. (2003, May). A general framework for clustering high-dimensional datasets. In Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on (Vol. 2, pp. 1091-1094). IEEE..

Highlights and Contribution of Manuscript for Agriculture Field:

Farm residences and greenhouses have the largest effect on reducing the soil surface for water infiltration. This decrease in area has negatively influenced the ecosystem heath of the region, as well as, decreasing the amount of agricultural land for soil-based agriculture and both surface and groundwater dynamics. Prime agricultural land is commonly located adjacent to urban centers, as historic human settlement and urbanization was associated with local land and water resources that could provide requisites of food and shelter. Urbanization poses an ongoing threat to farmland; every year an estimated 65,000 km2 of cropland is lost worldwide to urban expansion for uses such as housing, industry and infrastructure. At a time when there is growing pressure for local food security, prime agricultural lands in the peri-urban environment are becoming "sealed" from food production by the impervious surfaces. Global estimations indicate 0.43% of the world's land area is covered with impervious surfaces. Arable agricultural land is a scarce resource in Canada, as only 5% of it is suitable for crop production and free from limitations such as climate, topography, water availability and soil quality.