A Novel Approach for Weighted Clustering

CHANDRA B.
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi, India 110 016.
Email: bchandra104@yahoo.co.in

Abstract: - In majority of the real life datasets, all attributes do not have equal importance thereby same weights cannot be assigned for each attribute. Parametric Minkowski model [3] is used for considering the weightage scheme in the K-means clustering algorithm. There is always a challenge to obtain weights for the attributes from the data itself without pre-assigning weights. In this paper, a novel approach for automatic way of finding weights for various features in patterns has been proposed to address the problem of weights selection or weights generation in Parametric Minkowski model. The approach will play an important role for wider variety of clustering problems especially, where the attributes in a dataset do not have equal weights. The performance of the proposed approach has been evaluated over a number of datasets of UCI machine learning repository and compared with the performance of other popular clustering algorithms.

Key-Words: - Clustering, Parametric Minkowski Mode, Hyperlink-Induced Topic Search (HITS) Algorithm.

1 Introduction
Clustering [2, 11] an unsupervised technique is the process of organizing objects into groups such that similarity within the same cluster is maximized and similarities among different clusters are minimized. In many real world problems, clustering with equal weights for each attribute does not provide the desired results since different attributes have different significance levels.

Same weights are assigned to all the attributes in many clustering algorithms irrespective of the fact that all attributes do not have equal importance or weights in most of the real world problems. Weighted k-means clustering is considered as the popular solution to handle such kind of problems. In order to introduce the different weights for different attributes, parametric Minkowski model [3] is used to consider the weightage scheme in weighted k-means clustering algorithm. In parametric Minkowski model, the distance function is defined by a weighted version of the Minkowski distance measure. The parameters for this model are the weights in different dimensions. It is extremely difficult to determine the most suitable weights in the weighted k-means clustering since a number of complex mathematical equations are to be solved in parametric Minkowski model. Al-Harbi et. [16] al. has applied Simulated Annealing to generate weights for weighted k-means and Ayan et. al. [12] has used information gain as attribute weights but these approaches have inherent drawbacks of Simulated Annealing i.e. local optimization problem and drawbacks of information gain concept. In order to overcome the limitation of existing approaches, a novel approach based on Hyperlink-Induced Topic Search (HITS) algorithm [7, 8] has been proposed in this paper to automatically generate weights for the attributes in weighted k-means clustering algorithms. The proposed approach has been tested over a number of databases taken from UCI machine learning repository [4]. The proposed approach has also been compared with metric pairwise constraint k-means (MPCK-Means) [10], one of the most popular semi-supervised clustering algorithms.

Section 2 of the paper gives the overview of the work done in the area of weighted clustering and semi-supervised clustering along with the limitations of those algorithms. The proposed approach, parametric Minkowski model [3] and HITS algorithm [7] has been described in section 3. Results and discussions are shown in section 4 of the paper. Section 5 gives the comparative evaluation of the proposed approach with the existing method. Concluding remarks is given in the last section of the paper.

2 Related Work
In this section, K-means clustering algorithm and some of the widely known weighted clustering
algorithms have been described in brief with their limitations.

K-means [6] is one of the most popular clustering algorithms. K-means is a partitioning method, which creates initial partitioning and then uses iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The algorithm is used to classify a given data set into fixed number of clusters (K). K-means uses the concept of centroid, where a centroid represents the center of a cluster. In the process, K centroids, one for each cluster is defined apriori. Then each object of the data set is assigned a group with the closest centroid. The positions of k centroids are recomputed when all objects have been assigned to any of the clusters. The process is repeated until the centroids are no longer move. One of the limitations of K-means is not differentiating among attributes i.e. each attribute is given same importance in the clustering process. To overcome the limitation, weighted K-means is used with different weights for different attributes.

Al-Harbi et.al. [16] applied simulated annealing to generate weights for weighted K-means. Local optimization problem is one of the problems of simulated annealing techniques and that forms a major limitation in this approach too. Ayan et. al. [12] used information gain as attribute weights but the approach has the inherent drawbacks of the concept of information gain. A genetic cluster algorithm is also proposed by Demiriz et. al. [1] that has difficulty in defining fittest function as required in genetic process. In many real world problems, clustering in stand-alone mode does not provide the desired results. Semi-supervised clustering [5, 9-10, 15] is becoming popular with the presence of both labeled and unlabeled data in many practical problems.

Semi-supervised clustering uses a small amount of labeled objects (where information about the groups is available) to improve unsupervised clustering algorithms. Existing algorithms for semi-supervised clustering can be broadly categorized into constraint-based and distance-based semi-supervised clustering methods. Constraint-based methods [9, 10, 15] are generally based on pair-wise constraints i.e. pairs of objects labeled as belonging to same or different clusters to facilitate the algorithm towards a more appropriate partitioning of data. In this category, the objective function for evaluating clustering is modified such that the method satisfies constraints during the clustering process. In distance-based approaches [5, 10], an existing clustering algorithm uses a particular distance measure.

Wagstaff et. al. [9] has developed another variant of the k-means algorithm i.e. COP-Kmeans by incorporating background knowledge in the form of instance-level constraints. These instance-level constraints help in identifying which objects should be grouped together. An if-statement is introduced to assign cluster and ensures that none of the constraints are violated when the k-means algorithm groups each object to its closest cluster. However, the major limitation of this algorithm is that it does not allow violation of constraints even if it leads to a more cohesive clustering and leaving them vulnerable to noisy supervision. In order to overcome this limitation, Basu et. al. [15] has proposed pair-wise constraint k-means (PC-Kmeans) algorithm.

PC-Kmeans algorithm is similar to COP-Kmeans, but the main difference is that this algorithm can violate the constraints with some trade off as penalty for doing so. It tries to come up with a good cluster formation while minimizing the penalty that it incurs. A major limitation of this approach is that it assumes a single metric for all clusters, preventing them from having different shapes. Bilenko et. al. [10] has proposed metric pair-wise constraint k-means (MPCK-Means) algorithm to get rid of this limitation. MPCK-Means is considered as one of the most popular semi-supervised clustering algorithms in the recent past. Therefore, the proposed approach has been compared with MPCK-Means in the paper. The proposed approach based on Hyperlink-Induced Topic Search (HITS) algorithm is introduced to overcome the limitations of earlier work. Using the proposed approach the weights for the attributes are generated automatically from the data, for the weighted k-means using parametric Minkowski’s model, some of the preliminaries i.e. parametric Minkowski model and HITS algorithm are described in the next section of the paper.

3 Proposed Approach

Traditional clustering algorithms may not provide desired results for most of the real world problems, where all attributes do not have equally importance i.e. same weight cannot be assigned for each attribute. Parametric Minkowski model [3] is generally used to consider the weightage scheme in the clustering algorithm. In some of the earlier works, Mahalanobis distance has also been applied for capturing bias across different attributes. In
Mahalanobis distance, weights $\lambda_i$ of $ith$ attribute is the reciprocal of variance i.e. $\lambda_i = 1/\sigma_i^2$ whereas for parametric Minkowski model, weights $\lambda_i$ of $ith$ attribute represents the significance of that attribute which is not necessarily equal to the reciprocal of variance like most of the real world applications. It is also to be noted that Mahalanobis distance captures only bias of the attribute and not the significance of a particular attribute. Parametric Minkowski model is described as follows.

### 3.1 Parametric Minkowski Model

In parametric Minkowski model [3], the distance function is defined by a weighted version of the Minkowski distance measure. The parameters for this model are the weights in different dimensions. Let the weights for the $d$ dimensions are denoted by $\lambda_1, \ldots, \lambda_d$. For a pair of objects $X=x_1,\ldots,x_d$ and $Y=y_1,\ldots,y_d$, the parametric distance is defined as

$$ f(X,Y,\lambda_1,\ldots,\lambda_d) = \left( \sum_{i=1}^{d} \lambda_i \left\| x_i - y_i \right\|^p \right)^{1/p}. \quad (1) $$

A higher value of $\lambda_i$ indicates a greater significance for dimension $i$. In this model, the values of these weights are computed using p-norm mean square error function. The error function is also contained user defined similarity values for the different data points. A number of mathematical techniques such as gradient search method, linear search technique and bracket bisection techniques are further applied to solve the subsequent mathematical equations. Therefore, it is extremely difficult to determine the most suitable weights in the parametric Minkowski model. Hyperlink-Induced Topic Search (HITS) algorithm [7, 8] is used to automatically generate weights for attribute.

### 3.2 Hyperlink-Induced Topic Search

Hyperlink-Induced Topic Search (HITS) (also known as Hubs and authorities) [7] is a link analysis algorithm that rates Web pages It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. In the HITS algorithm, the first step is to retrieve the set of results to the search query. The computation is performed only on this result set, not across all Web pages. Authority and hub values (weights) are defined in terms of one another in a mutual recursion. An authority weight is computed as the sum of the scaled hub weights that point to that page. A hub weight is the sum of the scaled authority weights of the pages it points to. Relationship between hubs and authorities can be shown as a bipartite graph $G$, as in Fig. 1.

![Fig.1 Bipartite Graph, G representing Hubs and Authorities](image)

The algorithm performs series of iterations, each consisting of two basic steps:

**Authority Update:** Update each node’s Authority score (weight) to the sum of the Hub Score’s (weights) of each node that points to it. That is, a node is given a high authority weight by being linked to by pages that are recognized as Hubs for information.

**Hub Update:** Update each node’s Hub score (weight) to the sum of the Authority Score’s (weights) of each node that it points to. That is, a node is given a high hub weight by linking to nodes that are considered to be authorities on the subject.

$$ Hub_j = \sum_i Authority_i \& Authority_j = \sum_i Hub_j \quad \forall \ i, j \quad (2) $$

The Hub weight and Authority weight for a node is computed using HITS algorithm.

### 3.2.1 HITS Algorithm

The steps of the algorithm are given below

(a) Start with each node having a hub weight and authority weight of 1,

(b) Run the Authority Update Rule,

(c) Run the Hub Update Rule,

(d) Normalize the values by dividing each Hub weight by the sum of all Hub weights, and dividing each Authority weight by the sum of all Authority weights,

(e) Repeat from step (b) as necessary.

Since the hub and authority weights do not converge in the pseudo-code above, it is necessary to limit the number of steps that the algorithm runs for. One way to get around this, however, will be to normalize the hub and authority weights after each "step" by dividing each authority weight by the sum of all
authority weights, and dividing each hub weight by the sum of all hub weights. Sun et al [8] extended the HITS algorithm to Association Rule Mining by considering items as authorities and transactions as hubs.

### 3.3 Procedure

The steps for computation of weights for each numeric attribute \( A_i \) are as follows

(a) Discretization of numerical attribute using Equal Width Discretization (EWD)

(b) Transforming a numeric attribute into ten sequences of intervals and each interval is treated as a discrete value.

(c) Mapping of every numeric value into discrete value according to the interval it belongs to.

(d) Computation of weight for each discrete values \( w_{dj} \) and data objects \( (\text{Hub}_i) \) using HITS algorithm as shown in Figure 2 by considering data objects as hub and discrete values as authorities

(e) Finding weights for numeric attribute \( (W_{A_j}) \) using \( W_{A_j} = \sum \text{Hub}_i * w_{dj} \) where, \( i = 1,2, \ldots, \text{no. of data objects} \) and \( j=1,2, \ldots, \text{no. of attributes} \)

(f) Weighted k-means clustering algorithm is further applied to partition the dataset \( D \) into \( k \) partitions \( \{P_h\}_{h=1}^k \) using generated weights \( \lambda \) so that the objective function \( \Gamma_{\text{proposed}} \) is locally optimized. The objective function \( \Gamma_{\text{proposed}} \) is defined as

\[
\Gamma_{\text{proposed}} = \sum_{X_i \in X} f(X_i, C_{\gamma_i}, \lambda_1, \ldots, \lambda_d)
\]  

(3)

Where, \( f(X_i, C_{\gamma_i}, \lambda_1, \ldots, \lambda_d) \) is the parametric Minkowski distance between the data point \( X_i \) and its corresponding cluster centroid \( C_{\gamma_i} \) and is computed using equation (1).

### 4 Results and Discussion

Performance of the proposed approach has been illustrated on 12 well-known datasets taken from UCI machine learning repository [4]. It has been observed from the experimental results that the proposed approach outperforms the K-means algorithm. The details of datasets and results are given as follows.

**Iris**

Iris data set has four numeric attributes. The dataset contains 150 data objects of three different classes (Iris Setosa, Iris Versicolour and Iris Virginica). Significance of attributes (weight) for Iris data \([0.51, 1.00, 0.89, 0.98]\) is derived by proposed method discussed earlier. The second attribute is found to be the most significant and followed by the fourth attribute.

**Wpbc**

Wpbc data set contains 194 data objects described by 33 numeric attributes. The data objects belong to 2 different classes \([148, 46]\). Significance of attributes (weights) for wpbc data are \([0.23, 0.28, 0.37, 0.29, 0.31, 0.3, 0.31, 0.32, 0.38, 0.38, 0.36, 0.37, 0.37, 0.49, 0.66, 0.53, 0.45, 0.39, 0.48, 0.42, 0.29, 0.28, 0.29, 0.34, 0.38, 0.39, 0.33, 0.44, 0.45, 0.37, 1]\). It is found that 33th attribute is the most significant and followed by the 16th attribute.

**Sonar**

Sonar data set contains 208 data objects described by 60 numeric attributes. The data objects belong to 2 different classes \([97, 111]\). Significance of attributes (weights) for sonar data are \([0.59, 0.72, 0.83, \ldots, 0.23, 0.26, 0.27, 0.29, 0.34, 0.33, 0.41, 0.35, 0.33, 0.3, 0.3, 0.25, 0.24, 0.23, 0.26, 0.27, 0.31, 0.32, 0.34, 0.37, 0.34, 0.3, 0.27, 0.27, 0.26, 0.3, 0.27, 0.29, 0.28, 0.29, 0.37, 0.37, 0.38, 0.37, 0.35, 0.4, 0.51, 0.46, 0.51, 0.54, 0.41, 0.44, 0.5, 0.7, 0.64, 0.42, 0.38, 0.58, 0.55, 0.53, 0.63, 0.55, 0.79]\). It is found that 4th attribute is the most significant and followed by the 3rd attribute.

**Glass**

Glass data set contains 214 data objects described by 9 numeric attributes. The data objects belong to 6 different classes \([70, 76, 17, 13, 9, 29]\). Significance of attributes (weights) for glass data are \([0.47, 0.4, 0.38, 0.33, 0.39, 0.85, 0.49, 1, 0.64]\). It is found that 8th attribute is the most significant and followed by the 6th attribute.

**Thyroid**

Thyroid data set contains 215 data objects described by 5 numeric attributes. The data objects belong to 3 different classes \([150, 35, 30]\). Significance of attributes (weights) for thyroid data are \([0.25, 0.26, 0.51, 1, 0.88]\). It is found that 4th attribute is the most significant and followed by the 5th attribute.
attribute is the most significant and followed by the 5th attribute.

**Liver**

Liver data set contains 345 data objects described by 6 numeric attributes. The data objects belong to 2 different classes [145, 200]. Significance of attributes (weights) for liver data are [0.45, 0.34, 0.69, 0.61, 1, 0.64]. It is found that 5th attribute is the most.

**Ionosphere**

Ionosphere data set contains 351 data objects described by 34 numeric attributes. The data objects belong to 2 different classes [225, 126]. Significance of attributes (weights) for ionosphere data are [0.81, 1, 0.42, 0.24, 0.4, 0.23, 0.34, 0.22, 0.31, 0.24, 0.32, 0.2, 0.31, 0.19, 0.27, 0.21, 0.28, 0.18, 0.26, 0.19, 0.24, 0.18, 0.24, 0.17, 0.24, 0.18, 0.33, 0.17, 0.22, 0.18, 0.21, 0.17, 0.2, 0.2]. It is found that second attribute is the most significant and followed by the first attribute.

**Wdbc**

Wdbc data set contains 569 data objects described by 30 numeric attributes. The data objects belong to 2 different classes [357, 212]. Significance of attributes (weights) for wdbc data are [0.31, 0.26, 0.3, 0.42, 0.28, 0.31, 0.38, 0.31, 0.28, 0.3, 0.71, 0.34, 0.77, 1, 0.46, 0.39, 0.83, 0.4, 0.41, 0.69, 0.31, 0.22, 0.32, 0.43, 0.25, 0.33, 0.31, 0.2, 0.35, 0.39]. It is found that 14th attribute is the most significant and followed by the 17th attribute.

**Pima Indian Diabetes (PID)**

PID data set contains 768 data objects described by 8 numeric attributes. The data objects belong to 2 different classes [268, 500]. Significance of attributes (weight) for PID data are [0.41, 0.37, 0.49, 0.43, 1, 0.5, 0.58, 0.54]. It is found that 5th attribute is the most significant.

**Vowel**

Vowel data set contains 990 data objects described by 10 numeric attributes. The data objects belong to 11 different classes with 90 data objects of each class. Significance of attributes (weights) for vowel data are [0.73, 0.8, 0.87, 0.74, 0.86, 0.8, 1, 0.85, 0.73, 0.76]. It is found that 7th attribute is the most significant and followed by the 3rd attribute.

**Segmentation**

Image Segmentation data set contains 2310 data objects described by 19 numeric attributes. The data objects belong to 7 different classes with 330 data objects of each class. Significance of attributes (weights) for segmentation data are [0.1, 0.12, 1, 0.8, 0.94, 0.78, 0.99, 0.82, 0.99, 0.23, 0.26, 0.22, 0.24, 0.16, 0.14, 0.17, 0.2, 0.14, 0.45] is derived by proposed method discussed earlier. It is found that 3rd attribute is the most significant.

**Optdigit**

Optdigit data set contains 5620 data objects described by 64 numeric attributes. The data objects belong to 10 different classes [571, 557, 572, 568, 558, 558, 566, 554, 562, 554]. Significance of attributes (weights) for opdigit data are [1, 0.73, 0.15, 0.19, 0.19, 0.19, 0.67, 0.96, 1, 0.45, 0.18, 0.19, 0.15, 0.14, 0.52, 0.95, 1, 0.36, 0.16, 0.16, 0.17, 0.17, 0.49, 0.97, 1, 0.37, 0.17, 0.15, 0.21, 0.15, 0.46, 1, 1, 0.45, 0.18, 0.18, 0.21, 0.15, 0.32, 1, 0.98, 0.58, 0.24, 0.21, 0.17, 0.14, 0.32, 0.98, 1, 0.72, 0.13, 0.14, 0.14, 0.16, 0.32, 0.92, 1, 0.87, 0.15, 0.21, 0.21, 0.16, 0.55, 0.92]. It is found that 1st attribute are the most significant attribute.

5 Comparative Results

Performance of the proposed approach of finding weights for attributes using HITS algorithm in parametric Minkowski model was evaluated on large number of datasets taken from UCI machine learning repository and it was compared with the results of K-means, MPCK-Means (see Table 4). Rand Index [17and Jaccard coefficient [13] have been used as cluster validation measures. Index value has been considered on percentage scale by multiplying index with 100.

From Table 4, it is seen that the proposed approach performs better than K-means and MPCK-Means. Though, the proposed approach shows either equal or slightly improvement over some of the datasets like Wpbc, Liver etc. but significantly better than many benchmark datasets like Iris, Glass, Wdbc etc. It can be concluded from Table 1 that proposed approach performs better than K-means and MPCK-Means for most of the datasets.

6 Conclusions

In this paper, a novel approach based on HITS algorithm and parametric Minkowski model has been proposed in order to introduce weightage scheme and to generate the suitable weights in the clustering algorithm. The effectiveness of the proposed approach over the existing clustering algorithm has been illustrated using UCI machine
learning repository datasets and compared with the popular clustering algorithms such as K-Means and MPCK-Means. The proposed clustering approach produces better results even with the widely used semi-supervised clustering algorithm like MPCK-Means. It can be applied to large scale of practical problems as most of the real world problems do not have equal weights for each of the attributes and weights are either unknown or hard to obtain. The approach can play an important role for wider variety of clustering problems especially where the attributes of a dataset do not have equal weights.

References:
[12] F. Ayan, “Using information gain as feature weight,” in Proc. of 8th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN’99), Turkey 1999.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rand Index</th>
<th>Jaccard Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>71.82</td>
<td>89.30</td>
</tr>
<tr>
<td>Wpbc</td>
<td>53.35</td>
<td>51.49</td>
</tr>
<tr>
<td>Sonar</td>
<td>50.32</td>
<td>50.14</td>
</tr>
<tr>
<td>Glass</td>
<td>65.95</td>
<td>54.68</td>
</tr>
<tr>
<td>New-Thyroid</td>
<td>63.13</td>
<td>77.63</td>
</tr>
<tr>
<td>Liver</td>
<td>50.43</td>
<td>50.09</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>58.89</td>
<td>57.73</td>
</tr>
<tr>
<td>Wdbc</td>
<td>75.04</td>
<td>81.85</td>
</tr>
<tr>
<td>PID</td>
<td>55.07</td>
<td>51.13</td>
</tr>
<tr>
<td>Vowel</td>
<td>86.51</td>
<td>86.24</td>
</tr>
<tr>
<td>Segmentation</td>
<td>83.10</td>
<td>85.85</td>
</tr>
<tr>
<td>Optdigit</td>
<td>92.00</td>
<td>91.19</td>
</tr>
</tbody>
</table>

Table 1 Accuracy on UCI datasets