Chapter 5

Exploitation of Multivalued Type Proximity for Symbolic Feature Selection

5.0 Background

In this chapter, an important aspect of cluster analysis called curse of dimensionality is addressed, particularly for symbolic objects described by both interval and multivalued features. Proximity measures and clustering algorithms are computationally expensive if the number of features describing symbolic objects is very large. The length of feature vectors representing symbolic objects depends upon the number of features recorded and larger the length of the feature vectors, higher is the dimension of the feature space. Thus, besides creating analysis problem, it creates an added problem requiring large memory for storing symbolic objects. Moreover, it

* some part of the material in this chapter appeared in the following research papers:


is more likely that some features may be noisy or irrelevant and degrade the performance of learning / classification algorithms. Hence, removing the noisy features by keeping only the relevant features may lead to better and interpretable classification, reduction in time and increase in the comprehensibility of the results (Law et. al., 2004; Talavera, 1999 (a, b)). This operation is referred to as dimensionality reduction through feature selection. Selecting a subset of the original features though seems to be a simplest method of dimensionality reduction, selecting a best possible subset of features without requiring generation of all possible subsets is still a challenging issue. The term feature selection is referred to a process that outputs a subset of the feature set. The methods that create new features based on transformations or combinations of the original features are termed feature extraction algorithms (Jain and Douglas, 1997). A variety of feature transformation methods for dimensionality reduction have been proposed (Jain and Dubes, 1988; Nagabhushan, 1988; Nagabhushan and Sudhanva, 1989; Sudhanva and Gowda, 1992; Talavera, 200; Fodor, 2002; Mitra et.al., 2002; Jennifer, 2003; Law et. al., 2004; Huan, 2005).

This chapter proposes a method which is in covenant with the former type of dimensionality reduction. The proposed feature selection scheme exploits the multivalued type proximity measure proposed in chapter 2 and chapter 3. The proposed feature selection is applicable to unlabeled data. In literature, little work has been done on unsupervised feature selection scheme even for conventional data (Law et. al., 2004; Mitra et.al., 2002; Huan, 2005). Most of the feature selection algorithms reported in literature work only for supervised data where class information is already available (Jain and Douglas, 1997; Miller, 2002). But, to the best of our knowledge, no work has been reported in literature on unsupervised feature selection specifically for symbolic data. However, some feature extraction schemes (Nagabhushan et. al, 1995; Bock and Diday, 2002) based on principal component analysis (PCA) are available. But the PCA and other feature extraction algorithms tend to fail in identifying informative features if the objects are unlabeled (Wolf and Bileschi, 2005; Jennifer, 2003) and also it is difficult to get an intuitive understanding of data using only the extracted features.
In view of this, in this chapter a simple and efficient unsupervised feature selection scheme for symbolic data which works in the proximity space has been proposed. The proposed feature selection scheme is based on the Type 3 similarity / dissimilarity matrix (chapter 2 and 3). The proposed scheme has two levels of selection. The first level selection of features is achieved based on the standard deviation of the feature wise proximity values and subsequently, the selected subset is refined in the second level by adding back if any, the left out features based on the correlation / covariance of the feature wise proximity values. The proposed levels of feature selection work in the proximity space but not in the original feature space. Feature selection becomes possible in the proximity space as the perceived proximity value is of multivalued type. Experiments have been conducted on all four standard data sets and the results revealed that the proposed feature selection scheme is more effective in selecting significant features.

The chapter is organized as follows. Section 5.1 presents the stages in the proposed feature selection scheme. The results of the experiments conducted on the standard data sets are presented in section 5.2. Finally, conclusion is given in section 5.3.

5.1 Feature Selection: A Proposed Scheme

In this section, we exploit the multivalued proximity measures viz., symbolic similarity and symbolic dissimilarity measures proposed respectively in chapter 2 and chapter 3 for the purpose of reducing the dimensionality of the unlabeled symbolic objects through feature selection (subsetting). The proposed feature selection scheme consists of two stages. In the first stage, the more prominent features will be selected based on the standard deviation of the proximity values computed among all symbolic objects with respect to the features. Subsequently, the second stage focuses on refining the subset selected during the first stage by adding back some of the left out features. The features which are added back are expected to increase the cohesion of the clusters. Such features are selected based on correlation and covariance respectively if the reduction process is based on similarity matrix and dissimilarity matrix.
5.1.1 Feature Subset Selection

Let there be m number of objects say O₁, O₂, ..., Oₘ each is described by n symbolic features. The Type 3 proximity matrix (Type 3 similarity / dissimilarity matrix proposed in chapter 2/3) is computed. The computed proximity matrix is of size m x m with each element being of type multivalued of dimension n. The matrix M of size m² x n is constructed by listing out each multivalued type element of the Type 3 proximity matrix one by one in the form of rows. We recommend to compute the standard deviation for each column of the matrix M and also the average standard deviation due to all n columns.

Let Stdᵢ denote the standard deviation of the jᵗʰ column of the matrix M and let AvgStd denote the average standard deviation.

\[ \text{Std}_j = \text{Std}(M(i,j)) \quad \forall \ i = 1 \ldots m \]  
\[ \text{AvgStd} = \frac{\sum_{j=0}^{n} \text{Std}_j}{n} \quad \ldots (5.1) \]

It should be observed here that the computation of standard deviation is not in the feature space but in the proximity space. In fact computation of standard deviation cannot easily be applied on the original features as they are of either interval type or multivalued type. Based on the computed standard deviation, the features corresponding to the columns with their standard deviation greater than the average standard deviation are selected. That is, the feature Fⱼ is selected if \( \text{Std}_j > \text{AvgStd} \). The set of all such selected features defines the set of more prominent features. Since only the features which have their associated standard deviations more than the average standard deviation are selected, the selected features will have different proximity values among m objects and thus have better discriminating capabilities when compared to the other left out features.

5.1.2 Feature Subset Refinement

In this sub section, we propose a method of refining the subset selected through previous stage. It is felt that the above subset selection method, though reduces the
dimensionality of the symbolic objects it may not always generate a subset of all significant features and thus, some times it may remove the features which are more significant in increasing the cohesion of the clusters. In order to identify the features which are capable of increasing the cohesion among the objects in a cluster, in addition to selecting the features which are capable of discriminating the clusters, the following refinement process is recommended.

If the matrix $M$ is constructed by the use of the Type 3 similarity matrix, then we recommend to compute the sum of the total correlations of each column of $M$ with all the other columns. Among the features which are left out in the first stage, we select some of them which have their respective total correlation more than the average correlation and subsequently add them back. These new features which are getting added up are expected to increase the cohesion of the clusters of the objects.

Let $T_{Corr_j}$ be the total correlation of the $j^{th}$ column with all other columns of the matrix $M$ and let $AvgT_{Corr}$ be the average of all total correlation obtained due to $n$ columns.

\[ T_{Corr_j} = \sum_{k=0}^{n} Corr(j^{th} \text{ column, } k^{th} \text{ column}) \] \hspace{1cm} \ldots (5.2)

and

\[ AvgT_{Corr} = \frac{\sum_{j=0}^{n} T_{Corr_j}}{n} \] \hspace{1cm} \ldots (5.3)

If $T_{Corr_j} > AvgT_{Corr}$ then the feature $F_j$ corresponding to the $j^{th}$ column is selected to be the feature capable of increasing the cluster cohesion and thus it will be added back.

On the other hand, instead of using the similarity matrix, if the dissimilarity matrix is used then we recommend to take a decision based on the total covariance of columns in place of correlation.

The following is the detailed algorithm designed for the feature selection using symbolic similarity measure.
Algorithm 5.1: Feature Selection Scheme using similarity measure

Input: Set of m symbolic objects \( \{O_1, O_2, \ldots, O_m\} \) each is described by n symbolic features \( F_1, F_2, \ldots, F_n \) of dimension n.

Output: Dimensionally reduced symbolic objects.

Method:

Step 1. Compute the Type 3 similarity matrix and construct the matrix M of size \( m^2 \times n \) by listing out each multivalued type element of the Type 3 proximity matrix one by one in the form of rows.

Step 2. Calculate the standard deviation \( Std_j \) for each column of the matrix M and also the average standard deviation using eqn (5.1) due to all n columns. Select all features \( F_j \) with \( Std_j > AvgStd \).

Step 3. Compute the total correlation \( TCorr_j \) of the \( j^{th} \) column with all other columns of the matrix M and also the average total correlation using eqn (5.2) due to all n columns in matrix M. If \( TCorr_j > AvgTCorr \) (eqn (5.2)) and \( F_j \) is a left out feature in the Step 2 then add it back to the selected subset.

Step 3. Represent the symbolic objects with the feature subset selected.

Algorithm ends.

Instead of using the Type 3 similarity measure, if the Type 3 dissimilarity measure is used then we recommend to carry out step 3 of the algorithm based on covariance in place of correlation.

5.2 Experimental Results

In this section, we conduct a sequence of experiments to corroborate the success of the feature selection scheme proposed. For this purpose, we consider the same four standard data sets viz., fat oil (Appendix A), microcomputer (Appendix B), microprocessor (Appendix C) and city temperature (Appendix D). The proposed feature selection scheme has been experimentally validated on all four data sets by the use of Hubert T\(^*\) statistics (Jain and Dubes, 1988).
The MSV based clustering algorithm (algorithm 2.3) proposed in chapter 2, the MDV based clustering algorithm (algorithm 3.3) proposed in chapter 3 (for supervised analysis) and the agglomerative based similarity (algorithm 4.4) / dissimilarity (algorithm 4.3) approach (for unsupervised analysis) are used during experimentation. However, divisive based algorithms could also be used for validation purpose. For the computation of Hubert $\Gamma$ value with respect to an algorithm (chapter 2 / 3 / 4) we have used the clusters obtained by the algorithm using all the features with out any reduction in features as reference clusters.

In this section, we also justify the proposed principle of feature selection by giving $\Gamma$ values for various feature subsets generated by ranking the features based on the correlation / covariance of the respective feature similarity / dissimilarity values.

**Experiment 1**

In this experiment the proposed feature selection scheme is validated on the fat oil data set. The fat oil data set consists of $m = 8$ symbolic objects. The Type 3 similarity matrix of size $8 \times 8$ with each element being multivalued of dimension $n = 5$ is obtained through application of the similarity measure proposed in chapter 2. The obtained Type 3 similarity matrix is then converted into a matrix of size $64 \times 5$ by listing out the obtained multivalued similarity values one after the other in the form of rows.

Table 5.1 summarizes the standard deviation of each column of this transformed matrix. The standard deviation of the $j^{th}$ column represents the deviation in similarity values among all $m$ objects with respective to only the $j^{th}$ feature. As suggested, in the first stage (feature subset selection stage) of the proposed feature selection scheme, the features $F_1$, $F_2$ and $F_3$ are selected since their associated standard deviations are more than the average standard deviation. Thus $\{F_1, F_2, F_3\}$ is the subset selected initially.

In Table 5.1 we have also summarized the total correlation associated with each feature. As neither the feature $F_4$ nor the feature $F_5$ is associated with total correlation more than the average correlation, no feature is added back during the second stage of feature selection (feature subset refinement stage). Thus the final selected subset is
Table 5.1 Feature wise standard deviation and total correlation of the multivalued similarity on fat oil data

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>0.4527</td>
<td>0.3859</td>
<td>0.37657</td>
<td>0.2890</td>
<td>0.202255</td>
<td>0.3413</td>
</tr>
<tr>
<td>Correlation</td>
<td>2.7491</td>
<td>2.8925</td>
<td>3.1790</td>
<td>2.6186</td>
<td>2.7798</td>
<td>2.8438</td>
</tr>
</tbody>
</table>

Fig 5.1 Graph of Hubert $\Gamma$ statistics versus selected subset of features on fat oil data

(a) Hubert $\Gamma$ statistics computed for the clusters obtained through algorithm 2.3 on different subset of features selected with respect to the clusters obtained by the algorithm 2.3 considering all features.

(b) Hubert $\Gamma$ statistics computed for the clusters obtained through algorithm 4.4 on different subset of features selected with respect to the clusters obtained by the algorithm 4.4 considering all features.
\{F_1, F_2, F_3\}. Using only these 3 selected features the fat oil objects are clustered by the use of the algorithm 2.3 and the algorithm 4.4. The obtained clusters are then validated by the use of Hubert \( \Gamma \) statistics by considering the clusters obtained on fat oil data by the respective algorithm using all the five features as reference clusters. In case of the algorithm 2.3, the obtained \( \Gamma \) value is 1 and in case of algorithm 4.4 the obtained \( \Gamma \) statistics is 0.408.

Out of curiosity in knowing, if addition of any left out feature results with increasing \( \Gamma \) value, we have ranked the left out features \( F_4 \) and \( F_5 \) in the sequence \( F_5, F_4 \) based on the decreasing values of their associated total correlations. It shall be observed in Fig 5.1 that addition of \( F_5 \) does not increase the \( \Gamma \) value. Infact this is ruled out in case of algorithm 2.3 (Fig 5.1(a)) since the selected subset itself has the \( \Gamma \) value maximum 1. Even in case of algorithm 4.4 (Fig 5.1 (b)) the \( \Gamma \) value remains same justifying that the proposed feature selection scheme is effective in selecting the more appropriate features.

**Experiment 2**

In this experiment, we have considered another standard data set, the microcomputer data for validating the proposed feature selection scheme. In Table 5.2, the standard deviation and the total correlation associated with each feature of the data set is given. During the first stage of the proposed feature selection scheme, the features \( F_1, F_2 \) and \( F_4 \) are selected as their associated standard deviations are more-than the average standard deviation. During the second stage the feature \( F_3 \) is added back as its associated total correlation is more than the average correlation. Thus, finally the selected subset is \( \{F_1, F_2, F_3, F_4\} \). Using these selected features, the microcomputer objects are clustered by the use of the algorithm 2.3 and the algorithm 4.4. The obtained clusters are then validated by the use of Hubert \( \Gamma \) statistics by considering the clusters obtained on microcomputer data by the respective algorithm using all original five features as reference clusters. In case of both the algorithms the clusters obtained by considering the reduced feature set are identical to the expected clusters and thus both the obtained \( \Gamma \) values are maximum(1). Therefore, the proposed
Table 5.2 Feature-wise standard deviation and total correlation of the multi-valued similarity on microcomputer data

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>0.4969</td>
<td>0.4968</td>
<td>0.3965</td>
<td>0.4969</td>
<td>0.4345</td>
<td>0.4621</td>
</tr>
<tr>
<td>Correlation</td>
<td>2.3664</td>
<td>2.1123</td>
<td>2.6186</td>
<td>2.2741</td>
<td>1.8982</td>
<td>2.2577</td>
</tr>
</tbody>
</table>

Fig. 5.2 Graph of Hubert Γ statistics versus selected subset of features on microcomputer data

(a) Hubert Γ statistics computed for the clusters obtained by the algorithm 2.3 considering all features selected with respect to the clusters obtained by the algorithm 2.3 on different subset of features.

(b) Hubert Γ statistics computed for the clusters obtained by the algorithm 4.4 considering all features selected with respect to the clusters obtained by the algorithm 4.4 on different subset of features.
approach has good confining capability in figuring out the relevant features during feature selection.

Experiment 3

Another benchmark data set (microprocessor data) is also considered. Table 5.3 summarizes the standard deviation associated with each feature of the data set. As suggested, in the first stage of the proposed feature selection scheme, the features $F_2$, $F_4$ and $F_5$ are selected because, their associated standard deviations are more than the average standard deviation. In Table 5.3, we have also summarized the total correlation associated with each feature. In the second stage of the feature selection scheme, the feature $F_3$ is added back as its associated total correlation is more than the average correlation. Thus, finally the subset $\{F_2, F_3, F_4, F_5\}$ is selected. Using these selected features the microprocessor objects are clustered by the use of the algorithm 2.3 and the algorithm 4.4. The obtained clusters are then validated by the use of Hubert $\Gamma$ statistics by considering the clusters obtained on microprocessor data by the respective algorithm using all original five features as reference clusters. In case of the algorithm 2.3, the obtained $\Gamma$ value is 1 and in case of algorithm 4.4 the obtained $\Gamma$ statistics is 0.11.

Experiment 4

The proposed unsupervised feature selection scheme is applied on the city temperature data. Table 5.4 summarizes the standard deviation of each feature of the city temperature data. As suggested, in the first stage of the proposed feature selection scheme, the features from $f_4$ to $f_{10}$ are selected as their associated standard deviations are more than the average standard deviation. Thus $\{F_4, F_5, F_6, F_7, F_8, F_9, F_{10}\}$ is the subset selected initially.

In Table 5.4 we have also summarized the total correlation associated with each feature. As the features $F_3$ and $F_{11}$ are associated with total correlations more than the average correlation, they are added back during the second stage of the feature selection scheme. Thus, the final selected subset is $\{F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}\}$. Using these selected features the city temperature objects are clustered by the use of the algorithm 2.3 and the algorithm 4.4. The obtained clusters are then validated by
Table 5.3 Feature wise standard deviation and total correlation of the multivalued similarity on microprocessor data

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>0.4109</td>
<td>0.4859</td>
<td>0.3447</td>
<td>0.4410</td>
<td>0.4537</td>
<td>0.4272</td>
</tr>
<tr>
<td>Correlation</td>
<td>1.6127</td>
<td>1.5863</td>
<td>2.9415</td>
<td>2.6503</td>
<td>2.5205</td>
<td>2.2622</td>
</tr>
</tbody>
</table>

Fig 5.3 Graph of Hubert $\Gamma$ statistics versus selected subset of features on microprocessor

(a) Hubert $\Gamma$ statistics computed for the clusters obtained through algorithm 2.3 on different subset of features selected with respect to the clusters obtained by the algorithm 2.3 considering all features.

(b) Hubert $\Gamma$ statistics computed for the clusters obtained through algorithm 4.4 on different subset of features selected with respect to the clusters obtained by the algorithm 4.4 considering all features.
Table 5.4 Feature wise standard deviation and total correlation of the multivalued similarity on city temperature data

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>F11</th>
<th>F12</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>0.3503</td>
<td>0.3477</td>
<td>0.3622</td>
<td>0.3704</td>
<td>0.3702</td>
<td>0.3706</td>
<td>0.3653</td>
<td>0.3693</td>
<td>0.3797</td>
<td>0.3629</td>
<td>0.3522</td>
<td>0.3645</td>
<td></td>
</tr>
</tbody>
</table>

Fig 5.4 Graph of Hubert Γ statistics versus selected subset of features on city temperature data

(a) Hubert Γ statistics computed for the clusters obtained through algorithm 2.3 on different subset of features selected with respect to the clusters obtained by the algorithm 2.3 considering all features

(b) Hubert Γ statistics computed for the clusters obtained through algorithm 4.4 on different subset of features selected with respect to the clusters obtained by the algorithm 4.4 considering all features
the use of Hubert $\Gamma$ statistics by considering the clusters obtained on city temperature data by the respective algorithm using all the 12 features as reference clusters. In case of the algorithm 2.3 the obtained $\Gamma$ value is 0.797 and in case of algorithm 4.4 the obtained $\Gamma$ value is 0.809.

Further to know the effect, in case addition of any left out feature increases the $\Gamma$ value, the left out features are ranked in the sequence $F_2$, $F_{12}$, $F_1$ based on the decreasing value of their associated total correlation. It shall be observed in Fig 5.4 that the addition of the left out features in sequence does not increase the $\Gamma$ value and instead it deteriorates the $\Gamma$ value. Only when all the features are considered, the $\Gamma$ increases to maximum value. This has further revealed the capability of the proposed feature selection scheme in selecting the best subset of features without requiring any knowledge of the object class (unlabeled data). Table 5.5 summarizes the details on the features selected along with dimension for all the four data sets based on the Type 3 similarity matrix proposed in chapter 2. The Hubert $\Gamma$ statistics is also given in Table 5.5.

We have also conducted similar experiments on all the four data sets by the use of the Type 3 dissimilarity matrix proposed in chapter 3 and the complete experimental details are summarized in Table 5.6.

### 5.3 Conclusion

In this chapter, a novel feature selection scheme for reducing dimensionality of symbolic objects described by both interval and multivalued type features is proposed. The proposed feature selection scheme is based on the Type 3 similarity and dissimilarity matrices proposed in chapter 2 and chapter 3 respectively. In addition, the proposed selection scheme does not make use of the class labels of the data. The first stage of the feature selection algorithm remains common for both similarity and dissimilarity based schemes in which the standard deviations of the associated feature proximities are computed. However, in the second stage the algorithm computes the associated total correlation if Type 3 similarity measure is used otherwise, the total covariance if Type 3 dissimilarity measure is used. The proposed feature selection scheme has been experimentally validated on all four data sets by (Appendix A-D) the
Table 5.5 Hubert $\Gamma$ statistics for similarity based clusters obtained using supervised and unsupervised approaches

<table>
<thead>
<tr>
<th>Data set</th>
<th>Feature sub set selected and (dimension)</th>
<th>Hubert $\Gamma$ value (Supervised: Algorithm 2.3)</th>
<th>Hubert $\Gamma$ value (Unsupervised: Algorithm 4.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat oil (5 dimensional data)</td>
<td>$F_1$, $F_2$, $F_5$; (3)</td>
<td>1</td>
<td>0.408</td>
</tr>
<tr>
<td>Microcomputer (5 dimensional data)</td>
<td>$F_1$, $F_2$, $F_3$, $F_4$; (4)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Microprocessor (5 dimensional data)</td>
<td>$F_2$, $F_3$, $F_4$, $F_5$; (4)</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>City temperature (12 dimensional data)</td>
<td>$F_3$, $F_4$, $F_5$, $F_6$, $F_7$, $F_9$, $F_{10}$, $F_{11}$; (9)</td>
<td>0.797</td>
<td>0.809</td>
</tr>
</tbody>
</table>

Table 5.6 Hubert $\Gamma$ statistics for dissimilarity based clusters obtained using supervised and unsupervised approaches

<table>
<thead>
<tr>
<th>Data set</th>
<th>Feature sub set selected and (dimension)</th>
<th>Hubert $\Gamma$ value (supervised: Algorithm 3.3)</th>
<th>Hubert $\Gamma$ value (unsupervised: Algorithm 4.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat oil (5 dimensional data)</td>
<td>$F_1$, $F_2$, $F_5$; (3)</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Microcomputer (5 dimensional data)</td>
<td>$F_3$; (1)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Microprocessor (5 dimensional data)</td>
<td>$F_4$; (1)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>City temperature (12 dimensional data)</td>
<td>$F_1$, $F_2$, $F_8$, $F_9$, $F_{10}$, $F_{11}$, $F_{12}$; (7)</td>
<td>0.807</td>
<td>0.5</td>
</tr>
</tbody>
</table>
use of Hubert $\Gamma$ statistics (Jain and Dubes, 1988). For the computation of Hubert $\Gamma$ value we have used the clusters obtained by the algorithm (chapter 2 / 3 / 4) using all the features without any reduction as reference clusters. The obtained results have revealed that the proposed feature selection scheme has a very good capability in choosing more relevant features at a stretch without requiring generation of any combination of the features.