algorithm uses the heuristic search techniques for finding the suitable attributes. CLAS-Relief algorithm gives importance for redundancy in the feature sets. The novel CLAS-Relief algorithm is the extension of the LAS-Relief algorithm which is removing the irrelevant features as well as handling the redundant features in the feature space.

Irrelevant and redundant features are insignificant and removing them can enhance the learning process. Indeed, the feature selection process is alternatively viewed as process of identifying and removing as many irrelevant and redundant features as possible. CLAS-Relief algorithm combines characteristics of LAS-Relief algorithm and feature transformation techniques in order to achieve the efficient feature selection in the feature space.

Many research works were carried out in the past to overcome the noise issue in the feature space. Researchers are still at work in designing algorithms to reduce the noise issue. Hence it is necessary to develop new and more effective algorithms. Thus, a new feature selection algorithm is proposed to overcome the noisy issue.

6.2 Motivation and Objectives

A wide range of feature selection algorithms were developed in the past and those algorithms mainly focused on improving the classification accuracy. There are many factors which affect the feature selection algorithms. Among them, one of the important factors is noisy issue in feature space. Inferring the
valuable information from the high dimensional data with noise is a very tedious process even though it provides rich information. In some real world applications, the dataset consist of more noisy features that affect the classification accuracy. Extracting the valuable information with the noisy dataset is a challenging task in data mining. The following research works are carried out to improve the classification accuracy with the noisy nature of feature space.

Relief-C, an algorithm to overcome the noisy issue, was developed by Dash [Das, 2011]. In this algorithm, he implemented a high dimensional reduction technique in the Relief, which efficiently used for unlabelled data to select relevant features. He reported that Relief-C algorithm tackles the noisy issue in the feature space than other feature selection algorithms.

Robnik Sikonjam and Kononenko [Rob, 2003] proposed the new algorithm called ReliefF which was able to tackle the noisy issue of feature space. According to them, the ReliefF algorithm is able to handle noisy dataset in the multiclass classification.

Fan Wenbing et al., [Fan, 2012] proposed a new approach from Relief algorithm. This is called as an Adaptive Relief (A-Relief) algorithm. This algorithm mitigates the issue of Relief algorithm by dividing the instance set adaptively. In A-Relief algorithm each of the features is inspected deeply to detect the bogus feature before the feature was trained through A-Relief.
Heum Park and Hyuk-Chul Kwon [Heu, 2007] reported that new extended Relief algorithm showed better performances for all dataset than the Relief algorithm. He used new computational techniques to handle the noisy and incomplete dataset, which resulted in improved classification accuracy.

From the existing research works, it is observed that the noise in the feature space still remains as an issue. Thus, the objective is to propose a new feature selection algorithm to overcome the noisy issue in the feature space. The proposed algorithm addresses the issue of noise in the feature space to enhance the classification accuracy.

6.3 Novelty of CLAS-Relief Algorithm

The new feature selection called CLAS-Relief is proposed to address the noise issue in the feature space. This issue is handled in the proposed CLAS-Relief algorithm and its classification accuracy is enhanced. Here the classification accuracy is taken into account to ascertain the efficiency of the proposed algorithm. The proposed algorithm addresses the issue of noise and outlier in the feature space. In case of LAS-Relief feature selection algorithm, the outlier and noisy issue are not considered while selecting the features. The existence of noisy and outlier feature in the feature space reduces the classification accuracy in the data mining classification task. The proposed CLAS-Relief algorithm evolves from LAS-Relief algorithm. The basics of CLAS-Relief algorithm is discussed in the subsection below.
6.3.1 CLAS-Relief Algorithm

CLAS-Relief algorithm is the novel feature selection algorithm which follows the basics of LAS-Relief algorithm. This novel algorithm introduces the new routine for handling the noisy issue in the feature space in the original LAS-Relief algorithm. The Distance based Outlier Detection routine is used to tackle the noise issue in the feature space. This algorithm is named as CLAS-Relief algorithm. The schematic diagram of the CLAS-Relief algorithm is shown here in the Figure 6.1. The Figure 6.1 depicts the functionality of the feature selection.

Figure 6.1 Work flow of CLAS-Relief Algorithm

Figure 6.1 explains the schematic flow diagram of the CLAS-Relief algorithm. It has six steps to achieve the final feature set from the feature space. In the step 2, the Chebyschev distance based outlier detection routine is applied.
over the feature space to remove the noisy and outlier feature space. Thus step 3 gets noisy and outlier removed feature space. In the step 4, the proposed feature selection algorithm CLAS-Relief works for selection of highly relevant feature. Here, the Chebyshev distance is used along with LAS-Relief feature selection algorithm for feature weight estimation. The step 5 selects the highly relevant features above the user defined threshold value. Finally at step 6, the noisy and outlier features are removed feature space. The structure of CLAS-Relief algorithm and its components are elucidated below.

**CLAS-Relief Algorithm**

**Procedure: Distance based Outlier Detection**

**Step 1** Initializing the weight \([A] = 0.0\);

**Step 2** While \((T<=100)\)

**Begin**

**Step 2.1:** Select the instance \((X)\) by randomly

**Step 2.2:** Calculate the nearest Hit and nearest Miss

**Step 2.3:** \(W[i] = \text{Las-Median}(X)\)

**End**

**PROCEDURE : Las-Median(X)**

**Sub Las_Median(X)**

**Begin**

**Step 1:** Calculating the distance

\[
\text{Chebyshev distance}=\max(|p_i-q_i|)
\]
**Step 2: Sub Select(X,k)**

If |X| = 1 then return single Element in X

else

**Step 2.1:** if(|X| >= k) then

return Select (X,k)

else

if(|X_1| + |X_2| >=k) then

return a:

else

return Select (k+ |X_1|-|X_2|);

**Step 3:** While (i<= N)

Begin

W[i]=W[i] + diff(X(i),near miss(i)X )^2 - diff(X(i),near hit(i)X)^2

End

return (W[i]);

End

Where

[A] Attribute
T Threshold value
X Instances X
W[i] Weight of i th feature
\( K \) \hspace{1cm} \text{Smallest element in the weight vector}

\(|X_1| + |X_2|\) \hspace{1cm} \text{Sequence of Instances}

\text{near miss}(i) \hspace{1cm} \text{Near Miss value of i th instances}

\text{near hit}(i) \hspace{1cm} \text{Near Hit value of i th instances}

Distance based Outlier Detection routine is simple and efficient. This routine detects the noisy or outliers in the dataset based on the distance measure. An object in a dataset D is a distance-based outlier if at least a fraction \( \alpha \) of the objects in D is at a distance greater than r. The noisy free dataset are derived using this outlier detection routine. Then the noisy free dataset are taken for feature weight estimation for selection of feature above the threshold value.

6.4 Experiments

The efficiency of the proposed CLAS-Relief algorithm is measured in terms of classification accuracy. An experiment is conducted with the two different classifiers namely Naïve Bayes and J48, using two different dataset such as agriculture soil dataset and mushroom dataset obtained from the standard UCI repository. Firstly, the proposed CLAS-Relief algorithm is applied to dataset to select the relevant features. Secondly, the obtained relevant features are applied to Naive Bayes and J48 classifiers to calculate the classification accuracy. The results are discussed in the sub-section below.
6.4.1 Experiment with Agriculture Soil Dataset

The Agriculture soil dataset is a real time dataset collected from the agriculture research station. This dataset has been taken for the feature selection experiment for CLAS-Relief and LAS-Relief methods. The selected features are in turn used for classification, whether the Agriculture soil is fertile or non-fertile. The selected feature from the Agriculture soil dataset are taken to the classifier for accuracy estimation.

| Table 6.1 Comparison of CLAS-Relief and LAS-Relief Feature Selection Algorithms on Agriculture Soil Dataset |
|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Feature selection Algorithm Name                | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 |
|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Selected Times                                  | 18 | 16 | 4  | 13 | 14 | 6  | 11 | 8  | 3  | 7   |
| Selected Probability (%)                        | 89 | 83 | 18 | 72 | 80 | 14 | 76 | 23 | 15 | 24  |
| Relevant Features                               | F1, F2, F4, F5, F7                             | F10, F1, F2, F9, F8                             |
|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| 6.4.1.1 Accuracy Estimation on Agriculture Soil Dataset in Naive Bayes Classifier |

CLAS-Relief algorithm’s accuracy is higher than LAS-Relief algorithm. In the classification accuracy measure, Naive Bayes classifier is used for evaluating the feature selection algorithm. Precision, Recall and F measure are the vital parameters considered for the evaluation. Highly relevant features are
selected in CLAS-Relief algorithm. This results in enhanced accuracy. The results are depicted in the Table 6.2, where the Precision, Recall and F-measure values of CLAS-Relief algorithm are higher than the LAS-Relief algorithm.

Table 6.2 Accuracy of Results between CLAS-Relief and LAS-Relief in Naive Bayes Classifier on Agriculture Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS-Relief</td>
<td>0.765</td>
<td>0.768</td>
<td>0.762</td>
</tr>
<tr>
<td>LAS-Relief</td>
<td>0.753</td>
<td>0.756</td>
<td>0.754</td>
</tr>
</tbody>
</table>

The higher accuracy is achieved in the CLAS-Relief algorithm is due to the selection of highly relevant features from the feature space as well as tackling the noisy features in the feature space.

CLAS-Relief feature selection algorithm achieves higher accuracy on the Agriculture Soil Dataset in Naive Bayes classifier than LAS-Relief Feature Selection Algorithm. Apart from the accuracy in the classification, the CLAS-Relief algorithm tackles the noisy nature of feature space.
Figure 6.2 shows the enhanced classification accuracy in CLAS-Relief algorithm than LAS-Relief algorithm. From the Table 6.2 and Figure 6.2, it is clearly indicated that the performance of CLAS-Relief algorithm is higher than LAS-Relief algorithm.

**6.4.1.2 Accuracy Estimation on Agriculture Soil Dataset in J48 classifier**

The classification accuracy of the proposed CLAS-Relief algorithm was studied using J48 classifier. The results are compared with that of LAS-Relief, which are depicted in Table 6.3. The measures of comparison used were Precision, Recall and F-measure.

**Table 6.3 Accuracy of Results between CLAS-Relief and LAS-Relief in J48 Classifier on Agriculture Soil Dataset**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS-Relief</td>
<td>0.778</td>
<td>0.776</td>
<td>0.789</td>
</tr>
<tr>
<td>LAS-Relief</td>
<td>0.762</td>
<td>0.768</td>
<td>0.774</td>
</tr>
</tbody>
</table>

From the Table 6.3, it is understood that CLAS-Relief Feature Selection Algorithm shows higher accuracy in classification than LAS-Relief feature selection algorithm. The selection of highly relevant feature as well as handling the noisy nature of feature in the feature space by CLAS-Relief algorithm enhances the classification accuracy than the CLAS-Relief algorithm. The accuracy values of two methods are depicted in the Figure 6.3.
Figure 6.3 Accuracy of Results between CLAS-Relief and LAS-Relief in J48 Classifier for Agriculture Soil Dataset

From Figure 6.3, it is understood that the CLAS-Relief algorithm achieves the enhanced classification accuracy than LAS-Relief algorithm. It is because of the noisy and outlier removal in the feature space at the time of feature selection.

6.4.2 Experiment with Mushroom Dataset

The Mushroom Dataset is taken for feature selection experiments. In this experiment, the CLAS-Relief algorithm selects the more relevant features than LAS-Relief algorithm. These highly relevant features ultimately enhance the classification accuracy in the soil classification. The results indicate that CLAS-Relief algorithm selects the highly relevant features in the feature space and ignores the noisy and outlier features.
Table 6.4 Comparison of CLAS-Relief and LAS-Relief Feature Selection Algorithms on Mushroom Soil Dataset

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLAS-Relief</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selected Times</td>
<td></td>
<td>17</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>16</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Selected Probability (%)</td>
<td></td>
<td>90</td>
<td>11</td>
<td>21</td>
<td>35</td>
<td>85</td>
<td>79</td>
<td>74</td>
<td>86</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Relevant Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAS-Relief</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selected Times</td>
<td></td>
<td>18</td>
<td>15</td>
<td>8</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Selected Probability (%)</td>
<td></td>
<td>88</td>
<td>84</td>
<td>31</td>
<td>29</td>
<td>66</td>
<td>18</td>
<td>20</td>
<td>81</td>
<td>72</td>
<td>54</td>
</tr>
<tr>
<td>Relevant Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5 shows the selected probability of features as well as number of time selected. Though the feature F1 is selected in both of these algorithms, the selected probability of particular feature is higher in CLAS-Relief algorithm when compared to LAS-Relief algorithm.

6.4.2.1 Accuracy Estimation on Mushroom Dataset in Naive Bayes Classifier

The Mushroom dataset have been taken for accuracy estimation. These dataset are pre-processed before classification task by Naive Bayes classifier. These dataset are applied on two feature selection algorithm such as CLAS-Relief and LAS-Relief. After pre-processing of feature selection, these dataset are fed into the Naive Bayes classifier to estimate the classification accuracy. For the classification accuracy, Precision, Recall and F-measure are taken for evaluating the two algorithms. The values obtained after the classification process are tabulated in Table 6.6. Table 6.6 and Figure 6.4 clearly depict that CLAS-Relief algorithm
performs better when compared to LAS-Relief algorithm in the accuracy measure. CLAS-Relief algorithm considers all the relevant features and eliminates the irrelevant features from the feature space than LAS-Relief algorithm.

**Table 6.5 Accuracy of Results between CLAS-Relief and LAS-Relief in Naive Bayes Classifier on Mushroom Dataset**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS-Relief</td>
<td>0.786</td>
<td>0.788</td>
<td>0.764</td>
</tr>
<tr>
<td>LAS-Relief</td>
<td>0.779</td>
<td>0.781</td>
<td>0.748</td>
</tr>
</tbody>
</table>

From the Table 6.6, the Precision, Recall and F Measure values are higher in CLAS-Relief algorithm compared to LAS-Relief algorithm.

**Figure 6.4 Accuracy of Results between CLAS-Relief and LAS-Relief in Naive Bayes Classifier for Mushroom Dataset**

Figure 6.4 shows that CLAS-Relief algorithm gets enhanced accuracy over the LAS-Relief algorithm.
6.4.2.2 Accuracy Estimation on Mushroom Dataset in J48 Classifier

The same Mushroom dataset are fed into J48 classifier for classification. The classification accuracy in J48 classifier is measured by Precision, Recall and F-measure. CLAS-Relief algorithm ranks higher than that of LAS-Relief in the classification. CLAS-Relief algorithm selects the relevant features and thereby increases the accuracy of classification.

Table 6.6 Accuracy of Results between CLAS-Relief and LAS-Relief in J48 Classifier on Mushroom Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS-Relief</td>
<td>0.776</td>
<td>0.778</td>
<td>0.777</td>
</tr>
<tr>
<td>LAS-Relief</td>
<td>0.764</td>
<td>0.768</td>
<td>0.768</td>
</tr>
</tbody>
</table>

The values in Table 6.7 depict that CLAS-Relief algorithm holds higher accuracy than LAS-Relief algorithm.

Figure 6.5 Accuracy of Results between CLAS-Relief and LAS-Relief in J48 Classifier for Mushroom Dataset
From Table 6.7 and Figure 6.5, it is clear that the selection of highly relevant features in the feature sets is enough for classification instead of selecting all the features. This reduces the computation time and cost in classifying the dataset. This factor attributes to the efficiency of CLAS-Relief algorithm.

6.5 Chapter Summary

Relief algorithm is a simple and efficient on feature selection. This algorithm fails to tackle the noisy nature of feature space. LAS-Relief is one of the variants of Relief, which lacks in handling the noisy features in the feature space. Feature selection in the noisy feature space is discussed in this chapter. In this chapter, a new CLAS-Relief algorithm has been proposed to overcome the noisy issue in the feature space based on Chebyshev distance measure. CLAS-Relief selects the most relevant features in the noisy feature space. The efficiency of the CLAS-Relief is proved by conducting an experiment with two different classifiers namely Naive Bayes and J48. The agricultural dataset and Mushroom dataset were used for the experiments. The experiments revealed that CLAS-Relief efficiently selects the most relevant features in the feature space and shows that the classification accuracy is improved.

Having studied the proposed algorithms, LAS-Relief, ELAS-Relief, and CLAS-Relief, the following chapter summarizes the features of the proposed algorithms and also provides directions for further research work.