

Comparison of Feature Dimension Reduction Approach for Writer Verification

Rimashadira Ramlee, Azah Kamilah Muda, Nurul Akmar Emran

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
rimashadira@gmail.com; azah@utem.edu.my;
nurulakmar@utem.edu.my;

Abstract. Dimension reduction is useful approach in data analysis application. In this paper, research is done to test whether the concept of Dimension reduction can be applied to improve writer verification process results. Two approaches have been chosen to be compared which are Features Selection and Feature Transformation, where the comparison is on the way of reducing the dimension of writer handwritten data. Both approaches have slightly difference results in reducing the data and classification accuracy. The objective of this paper is to observe the differences between both approaches according to the classification accuracy results, by using some classification techniques.

Keywords: Dimension Reduction, Writer Verification, Features Transformation, Features Selection

1 Introduction

Dimension reduction (DR) is a useful approach to solve a problem in data analysis application. Usually DR can be beneficial not only for reasons of computational efficiency but also because it can improve the accuracy of the analysis [1]. Reduction of the data dimension will help the process of identifying the most important features in handwritten data. According to [2], the process is either by transforming the existing features to a new reduced set of features or by selecting a subset of the existing features. In the data analysis, not all the features can yield important information that represents unique individualities of the writer, because maybe there is a lot of data redundancy which is not very useable in the analysis. In these issues, dimension reduction is useful to in order to improve that quality of the data used in analysis of data.

The purpose of this paper is to observe the comparison between features selection and features transformation approach in acquiring the most significant features among handwritten data via DR concept. The Comparison will be conducted by examining the classification accuracy and number of features data has been effectively reduced using both methods above. Features selection will select the feature directly from the original features and wish to keep the original meaning of the features, where Feature Transformation will allow the modification of the feature to a new feature space and

wish to determine which of those important features [3]. This paper is organized as follows in section 2 the detail explanation of Writer Verification is provided. In section 3 the description about Dimension Reduction approach and the way of their performing will be showed. Section 4 will describe about the Experiment setup of the process. The experiment illustration and result explanation will be in Section 5. Finally, conclusion will be in section 6.

2 Writer Verification

In theory, Writer Identification and Writer Verification belong to the group of behavioral methods in biometrics. Both methods will come to a conclusion of identifying the unknown writer, but the difference is according to the task of their performance. Writer Verification task is determined whether two samples of handwriting is written by the same writer or not [4].

Most of the recent research focuses on signature verification especially in field of on-line writer verification, where the verification process is used to perform the matching of two sample signature from one writer. To solve the problem of forged handwriting, dynamic information such as velocity, acceleration, and force exerted on the pen are utilized [5]. In this research, the verification process is chosen to be performed in text verification, because this task consists in matching the unknown writer with each of those in the selected subset.



Fig. 1. Example of Verification Process

Based on Fig 1, Writer verification also can be defined as one-to-one comparison process to make a decision for determine the real writer of handwritten document [5]. According to Srihari, Arora and lee, the individuality of handwriting rests on the hypothesis that each individual has consistent handwriting that is distinct from the handwriting of another individual.

3 Dimension Reduction Method

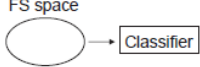
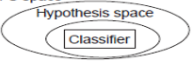
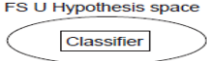
3.1 Feature Selection

Feature Selection is a popular approach used in Handwriting Analysis research field. In general, feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Feature selection techniques can be organized into three categories as showed in Table 1, depending on how they combine the feature selection search with the construction of the classification model which are filter methods, wrapper methods, and embedded methods [6].

In this work, we focus on filter model in our experimental framework due to its computational efficiency and usually chosen when the number of features becomes very large [7]. Different from wrapper method and embedded method, the classifier

chosen has to embed in the features selection approaches, where in this experiment features selection and classification process are separated. In further review, filter model can be categorized into two groups namely, Feature Weighting Algorithm (FWA) and Subset Search Algorithm (SSA).

Table 1. Type of Feature Selection Methods

	<p><i>Filter methods:</i> assess the relevance of features by looking only at the intrinsic properties of the data.</p>
	<p><i>Wrapper methods:</i> embed the model hypothesis search within the feature subset search.</p>
	<p><i>Embedded methods:</i> the search for an optimal subset of feature is built into the classification construction</p>

Additionally, the filter model relies on general characteristic of the training data to select some features without involving any learning algorithm. This will bring us to the objective of features selection to obtain the most important features by avoiding over fitting and improve model performance beside of provide faster and more cost-effective models, and gain a deeper insight into the underlying processes that generated the data [7]. There are three methods from filter model that have been selected:

- **Correlation-based Feature Subset Selection (CFS)**

CFS is a fully automatic algorithm, this method does not required user to specify any thresholds or the number of features to be selected, although both are simple to incorporate if desired [8]. The features subset evaluation function is:

$$M_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{K+K(K-1)\overline{r_{ii}}}} \tag{1}$$

Where M_{zc} is the correlation between the summed features and the outside variable, K is the number of variables, $\overline{r_{zi}}$ is the average of the correlations between the components and the outside variable, and $\overline{r_{ii}}$ is the average inter-correlation between features.

- **Relief**

Relief is a feature weighting algorithm that is sensitive to feature interactions. A key idea of the original Relief Algorithm is introduced by Kira and Rendell [9]. Relief will search for its two nearest neighbors, one is from the same class called *nearest hit*, and the other is from different class called *nearest miss*. Below is a probability for the weight of a feature:

$$W_{feature} = P(X|Y \text{ of different class}) - P(X|Y \text{ of same class}) \tag{2}$$

Where $W_{feature}$ is a weight, X is feature's difference value, and Y is a nearest instance.

- **Fast Correlation-based Filter (FCBF)**

In this algorithm, Symmetrical Uncertainty calculates dependency of features and finds best subset using backward selection technique with sequential search strategy [10]. FCBF can remove a large number of features that are redundant peers with predominant feature in the current iteration. Concept of entropy is used in this algorithm, for example, the entropy of a variable X is defined as:

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i)) \quad (3)$$

And the entropy of X after observing values of another variable Y is defined as:

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \quad (4)$$

Symmetrical uncertainty (SU) compensates for information gain's bias toward attributes with more values and normalizes its value to the range [0, 1]. Below is the equation for calculating symmetrical uncertainty coefficient:

$$SU(X, Y) = 2 \left[\frac{IG(X|Y)}{H(X) + H(Y)} \right] \quad (5)$$

An SU value of 1 indicates that using one feature compared to other feature's value can be totally predictable and value 0 indicates two features that are totally independent [13].

3.2 Feature Transformation using Principal Component analysis

Feature transformation refers to a family of data pre-processing techniques that transforms the original features of a data set to an alternative, a more compact set of dimensions, while retaining as much information as possible [3]. This techniques aim to reduce the dimensionality of data to a small number of dimensions which are linear or non-linear combinations of the vector coordinates in the original dimensions [11]. Principal Component Analysis (PCA) which is one of the unsupervised feature transformation techniques. Unsupervised technique does not take class labels into account so that the process became easier. Since PCA is a powerful tool for analyzing and identifying a valuable pattern in the data, we propose this technique in one of pattern recognition field which is handwriting analysis. Once the pattern of the data is found this technique will reduce the dimension without losing many features components from the original data like stated in [12].

Typically in this work, the objective of PCA is to transform the data into another set of feature f' , for example x_i transformed into x'_i in k dimensions shows:

$$x'_i = Wx_i \quad (6)$$

The transformation of PCA is by reducing the space that captures most of the variance in the data. The whole idea of PCA is rest on the covariance matrix of the data as:

$$C = \frac{1}{n-1} XX^T \quad (7)$$

C Captures the variance in the individual features and the off-diagonal terms quantify the covariance between the corresponding pairs of features. C can produce C_{PCA} , when the data is transformed by $Y = PX$ where the rows of P are the eigenvector of XX^T , then

$$C_{PCA} = \frac{1}{n-1} YY^T \quad (8)$$

$$C_{PCA} = \frac{1}{n-1} (PX)(PX)^T \quad (9)$$

C_{PCA} , is the quantifier of variance of the data in the direction of the corresponding principal component.



4 Experiment Setup

In this experiment, three representative feature selections are chosen in comparison with PCA like stated above.

4.1 Dataset

Handwriting sample dataset is taken from IAM Handwriting Database [13] is chosen to be used, where there are 657 classes available, however only 60 classes are used for this research. From these 60 classes, 4400 instances are collected, and are randomly divided into five different datasets to form training and testing dataset. First of all, the form of handwriting text will be extracted by using United Moment Invariance (UMI) [14]. We use both undiscretized and discretized data in this experiment in order to see the influence of verification and to improve the classification accuracy. Discretization was done by employed Equal Width Binning (EWB), and the main goal of this process is to minimize the number of intervals without significant loss of class-attribute mutual dependence [15]. Besides that, discretization process is important in order to obtain the detachment of writer's individuality and produce better data representation. Table 2 shows the example of data after UMI process:

Table 2. Example of Handwriting Data

Word	F1	F2	F3	F4	F5	F6	F7	F8
	0.75	0.38	0.44	0.19	0.67	4.62	0.50	4.59
	0.71	0.37	0.49	0.23	0.73	4.66	0.63	4.13

4.2 Framework Design

We design our work following the traditional of pattern recognition task for writer verification process as shown in Fig 4 below:

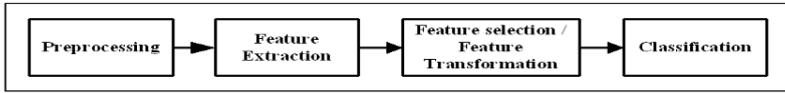


Fig. 2. Writer Verification Framework

This process begins, with preprocessing task, which is to process the data before extracting the real word features. UMI is applied in feature extraction part where all the handwriting text is changed to the word features representation. After discretization process the data becomes more clean and easy to determine the unique feature of the writer's data before we proceed to feature selection and classification task. We develop this experiment by using Waikato Environment for Knowledge Analysis (WEKA) 3.7.5 to measure the performance of feature selection, features transformation and classification method.

4.3 Verification Process

Verification will classify the data according to the same writer, where classification process will be used to verify the writer of document in sample data based on their class. There are several techniques of classification are used to perform the task which are Bayes Network, K-Nearest Neighbor (KNN), Random Forest and 1-Rule.

5 Result and Discussion

Two categories values are used to measure the performance of feature selection and feature transformation algorithm in writer verification process. The result as shown in Table 3:

Table 3. Result of the Experiment

5.1 Comparison by Classification Accuracy

The results have shown that when the data is undiscretized almost all the methods yield the same accuracy, which the average of all accuracies is less than 50%,

Sample Data	Category	Undiscretized				Discretized			
		CFS	Relief	FCBF	PCA	CFS	Relief	FCBF	PCA
Sample 1	Classification Accuracy	45.42%	45.99%	46.55%	45.54%	98.31%	98.08%	98.31%	66.33%
	Number of Selected features	1	8	8	5	6	8	8	6
Sample 2	Classification Accuracy	48.53%	45.99%	48.65%	49.24%	98.24%	98.24%	98.00%	65.22%
	Number of Selected features	1	8	8	5	6	8	8	6
Sample 3	Classification Accuracy	39.91%	45.99%	40.36%	40.02%	97.49%	97.38%	97.72%	62.14%
	Number of Selected features	1	8	8	5	6	8	8	6
Sample 4	Classification Accuracy	29.77%	45.99%	29.99%	30.43%	98.24%	97.91%	97.91%	55.57%
	Number of Selected features	1	8	8	5	5	8	8	6
Sample 5	Classification Accuracy	39.12%	45.99%	40.25%	39.46%	98.07%	97.96%	98.19%	61.90%
	Number of Selected features	1	8	8	5	6	8	8	6
AVERAGE		40.55%	45.99%	41.16%	40.94%	98.07%	97.91%	98.02%	62.23%

but the values are still slightly different as compared to the others. In comparison when the data has been discretized, the result plots better accuracy. The result of Relief method is slightly different among other feature selection methods, as compared to the chosen feature transformation method named PCA. Based on the observation, Relief method has gained higher average of classification accuracy when using undiscritized data, followed by FCBF, PCA and CFS. Even though the accuracy of relief method is decrease when the data is discretized but the different is not huge compared to others method. From the average of classification accuracy as shown in Table 3, Relief method is the best feature selection approach to verify the writer of sample data. It is caused by the process of feature selection that keeps the original meaning of every feature. Besides that, it is more helpful in verification process, rather than to modify the feature and transform it into a new feature space, So that the results will be more accurate when using this process.

5.2 Comparison by Selected Features

Second comparison is done by comparing the number of feature that can be reduced by each method from both approaches. Based on the result in TABLE III, The best method in reducing the features is CFS and PCA followed by Relief and FCBF. This because CFS has reduced 7 features and PCA has reduced 3 features using undiscritized data. Otherwise, only 2 features are reduced by both methods when the data is discretized, this because cleanliness of data can affect relation between the features. CFS and PCA can reduce the some feature according to their correlation of features between each other. CFS calculates the correlations and then searches the feature subset space which is important to represent the original. PCA also reduce the features but different concept from CFS which is less importance to represent the original data would reduced after transforming the feature by using all the original feature in data set. Relief and FCBF are not reducing any feature in both types of data. According to the Relief and FCBF concept, the relevant features are chosen rely on the dependences between each other. So that, this both methods will estimate that the entire feature are interact to each other in representing the original data then the reduction process cannot be done. Here, we able to prove that feature selection and features transformation approach successfully can reduce the dimension of data by only select the most significant feature that can verify the actual writer in verification process.

6 Conclusion

As a conclusion, the experimental result has proved that Dimension Reduction process can be used in verification process especially in processing data activities. Dimension reduction is more concern in eliminating the redundant data, so that this characteristic can improve the performance of the process. Redundancy will increase the relation among the feature and will cause the feature strongly depend on each other. CFS and PCA have different concept than Relief and FCBF as stated above. Among them CFS is the best method because, this method reduces the feature to the lowest number and accurately verify the writer of sample data.

Acknowledgement

This work is funded by the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM).

References

1. P.Cunningham, "Dimension Reduction", Technical Report UCD-CSI-2007-7, University College Dublin, 2007.
2. Ch. Aswani Kumar, "Analysis of unsupervised dimensionality reduction techniques", *Computer Science and Information Systems*, 6(2):217-227, doi: 10.2298/csis0902217K, 2009.
3. M.Masaeli, G.Fung, Jennifer G.Dy, "From Transformation-Based Dimensionality Reduction to Feature Selection", in Mahdokht Masaeli, G.F.J.G.D., ed. *Proc IEEE, 27th International Conference on Machine Learning*. Boston, USA, 2010.
4. S.Srihari, H.Arora, S.Lee, "Individuality of Handwriting", *Journal of Forensic Sciences*, 47(4), 2002, pp.1-17.
5. F.Leclerc, R.Plamondon: "Automatic signature verification and writer identification: The state of the art - 1989-1993", *Int. J. of Pattern Recogn. Artif. Intell. (IJPRAD)*, 8, 3, pp.643-660(1994).
6. Pratama, S.F., Muda, A.K., Choo, Y.-H.: Feature Selection Methods for Writer Identification: A Comparative Study. In: *Proceedings of 2010 Intl. Conference on Computer and Computational Intelligence*, pp. 234--239. IEEE Press, Washington (2010).
7. Saeys, Y., Inza, I., & Larranaga, P., "A Review of Feature Selection Techniques in Bioinformatics". *Journal of Bioinformatics*, 2507-2517, 2007.
8. M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", 1998, Hamilton, New Zealand.
9. M.Robnik-Sikonja, I.Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", *Machine Learning Journal*, 23-69, 2003.
10. L.Yu, H.Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", *Proc of the Twentieth International Conference on Machine Learning*, pp. 856-863, 2003.
11. T. Jolliffe, "Principal Component Analysis", Springer-Verlag, New York. 1986.
12. L. I. Smith, "A tutorial on Principal Components Analysis", PP. 13-27, February 26, 2002.
13. U.Marti, H. Bunke, "The IAM-database: an English Sentence Database for Off-line Handwriting Recognition", *International Journal on Document Analysis and Recognition*, Volume 5, 39-46, 2002.
14. S.Yinan, L. Weijun, W. Yuechao, "United Moment Invariants for Shape Discrimination", *International Conference on Robotics, Intelligent Systems and Signal Processing*, pp. 88-93, Changsha: IEEE, 2003.
15. Kotsiantis, S,K.&D., "Discretization Techniques: A Recent Survey", *GESTS International Transactions On Computer Science and Engineering*, 32, pp.47-58, 2006.