DOI: http://dx.doi.org/10.21609/jiki.v9i2.384

FEATURE SELECTION METHODS BASED ON MUTUAL INFORMATION FOR CLASSIFYING HETEROGENEOUS FEATURES

Ratri Enggar Pawening[1], Tio Darmawan[2], Rizqa Raaiqa Bintana[2,3], Agus Zainal Arifin[2] and Darlis Herumurti[2]

[1]Department of Informatics, STT Nurul Jadid Paiton, Jl. Pondok Pesantren Nurul Jadid Paiton, Probolinggo, 67291, Indonesia
[2]Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya, 60111, Indonesia
[3]Department of Informatics, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, Jl. H.R Soebrantas, Pekanbaru, 28293, Indonesia

E-mail: enggar.r@gmail.com[1], agusza@cs.its.ac.id[2]

Abstract

Datasets with heterogeneous features can affect feature selection results that are not appropriate because it is difficult to evaluate heterogeneous features concurrently. Feature transformation (FT) is another way to handle heterogeneous features subset selection. The results of transformation from non-numerical into numerical features may produce redundancy to the original numerical features. In this paper, we propose a method to select feature subset based on mutual information (MI) for classifying heterogeneous features. We use unsupervised feature transformation (UFT) methods and joint mutual information maximation (JMIM) methods. UFT methods is used to transform non-numerical features into numerical features. JMIM methods is used to select feature subset with a consideration of the class label. The transformed and the original features are combined entirely, then determine features subset by using JMIM methods, and classify them using support vector machine (SVM) algorithm. The classification accuracy are measured for any number of selected feature subset and compared between UFT-JMIM methods and Dummy-JMIM methods. The average classification accuracy for all experiments in this study that can be achieved by UFT-JMIM methods is about 84.47% and Dummy-JMIM methods is about 84.24%. This result shows that UFT-JMIM methods can minimize information loss between transformed and original features, and select feature subset to avoid redundant and irrelevant features.

Keywords: *Feature selection, Heterogeneous features, Joint mutual information maximation, Support vector machine, Unsupervised feature transformation*

Abstrak

Dataset dengan fitur heterogen dapat mempengaruhi hasil seleksi fitur yang tidak tepat karena sulit untuk mengevaluasi fitur heterogen secara bersamaan. Transformasi fitur adalah cara untuk mengatasi seleksi subset fitur yang heterogen. Hasil transformasi fitur non-numerik menjadi numerik mungkin menghasilkan redundansi terhadap fitur numerik original. Dalam tulisan ini, peneliti mengusulkan sebuah metode untuk seleksi subset fitur berdasarkan *mutual information* (MI) untuk klasifikasi fitur heterogen. Peneliti menggunakan metode *unsupervised feature transformation* (UFT) dan metode *joint mutual information maximation* (JMIM). Metode UFT digunakan untuk transformasi fitur non-numerik menjadi fitur numerik. Metode JMIM digunakan untuk seleksi subset fitur dengan pertimbangan label kelas. Fitur hasil transformasi dan fitur original disatukan seluruhnya, kemudian menentukan subset fitur menggunakan metode JMIM, dan melakukan klasifikasi terhadap subset fitur tersebut menggunakan algoritma *support vector machine* (SVM). Akurasi klasifikasi diukur untuk sejumlah subset fitur terpilih dan dibandingkan antara metode UFT-JMIM dan Dummy-JMIM. Akurasi klasifikasi rata-rata dari keseluruhan percobaan yang dapat dicapai oleh metode UFT-JMIM sekitar 84.47% dan metode Dummy-JMIM sekitar 84.24%. Hasil ini menunjukkan bahwa metode UFT-JMIM dapat meminimalkan informasi yang hilang diantara fitur hasil transformasi dan fitur original, dan menyeleksi subset fitur untuk menghindari fitur redundansi dan tidak relevan.

Kata Kunci: *Fitur heterogen, Joint mutual information maximation, Seleksi fitur, Support vector machine, Unsupervised feature transformation*

1. Introduction

Data and features which have high-dimensional are the main problems in the classification of supervised and unsupervised learning, which is becoming even more important with the recent explosion of the size of the available datasets both in terms of the number of data samples and the number of features in each sample. The rapid training time and the enhancement of classification accuracy can be obtained when dimension of data and features are decreased as low as possible.

Dimensionality reduction can be conducted using feature extraction and feature selection methods. Feature extraction methods transform the original features into a new feature which has lower dimension. The common used methods are principal component analysis (PCA) [1-2] and linear discriminant analysis (LDA) [3-4]. Feature selection methods is conducted by selecting some important features which minimises a cost function.

Feature selection methods are divided into two categories in terms of evaluation strategy, in particular, classifier dependent and classifier independent. Classifier dependent is divided into two methods, wrapper and embedded methods. Wrapper methods evaluate subsets of variables to detect the possible interactions between variables by measuring the prediction accuracy of a classifier. Wrapper methods had researched by [5-6]. They perform well because the selected subset is optimised for the classification algorithm. Wrapper methods may suffer from over-fitting to the learning algorithm and has very expensive in computational complexity, especially when handling extremely high-dimensional data. It means that each change of training models will decrease the function of subsets.

The feature selection stage in the embedded methods is combined with the learning stage [6]. Embedded methods perform variable selection as part of the learning procedure and are usually specific to given learning machines. These methods are less computational complexity and over-fitting. However, they are very specific and difficult for generalisation.

Classifier independent can be called as filter methods. Filter methods assess the relevance of features by looking only at the intrinsic properties of the data. The advantages of filter methods are: they can scale of high-dimensional datasets, they are computationally simple and fast, and they are independent of the classification algorithm. The disadvantage of filter methods is that they ignore the interaction between the features and the classifier (the search in the feature subset space is separated from the search in the hypothesis space), and most proposed techniques are univariate. Feature selection using filter methods is researched by [7]. These methods rank features according to their relevance to the class label in the supervised learning. The relevance score is calculated using mutual information (MI).

Information theory has been widely applied in filter methods, where information measures such as mutual information are used as a measure of the features's relevance and redundancy. MI can overcome problems of filter methods. Some methods which apply MI are MIFS [8], mRMR [9], NMIFS [10], and MIFS-ND [11]. These methods optimize the relationship between relevance and redundancy when selecting features. The problems of these methods is the overestimation of the significance of the feature candidates. The method for selecting the most relevant features using joint mutual information (JMI) is proposed by [12]. Joint Mutual Information Maximation (JMIM) is the development of JMI that adds ma-ximum of the minimum method.

Datasets with heterogeneous features can affect feature selection results that are not appropriate because it is difficult to evaluate heterogeneous features concurrently. Feature transformation (FT) is another way to handle heterogeneous features subset selection. FT methods unify the format of datasets and enable traditional feature selection algorithms to handle heterogeneous datasets. FT methods for heterogeneous features using unsupervised feature transformation (UFT) has proposed by [13]. The results of transformation from non-numerical into numerical features may produce redundancy to the original numerical features. The redundant features can be handled by selecting of the significant feature.

In this paper, we propose a method to select feature subset based on mutual information (MI) for classifying heterogeneous features. This paper is organized as follows. Section 2 describes research methodology of the proposed methods. Secti-

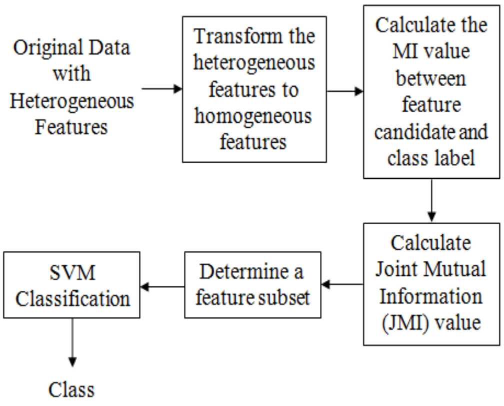
Original Data with Heterogeneous Features
Transform the heterogeneous features to homogeneous features
Calculate the MI value between feature candidate and class label
Calculate Joint Mutual Information (JMI) value
Determine a feature subset
SVM Classification
Class

Figure 1. The proposed methods

on 3 describes the conducted experiments and discusses the results. Section 4 concludes this study.

2. Methods

General description of the research methods is shown in Figure 1. The stages of UFT-JMIM methods in this study are transformation of heterogeneous features, calculation of MI value between feature candidate and class label, calculation of JMI value, determine a feature subset, and classification by using SVM.

Transformation of heterogeneous features

In the transformation stage, we transform the datasets that have heterogeneous features to homogeneous features. Feature transformation is conducted by using UFT methods. UFT is derived from the analytical relationship between MI and entropy. The purpose of UFT is to find a numerical *X'* to substitute the ori-ginal non-numerical feature *X*, and *X'* is constrai-ned by *I*(*X';X*) = *H*(*X*). This constraint makes the MI between the transformed *X'* and the original *X* to be the same as the entropy of the original *X*.

This condition is critical because it ensures that the original feature information is preserved, when non-numerical features are transformed into numerical features. It is also worth noting that the transformation is independent of class label, so that the bias introduced by class label can be reduced. After it is processed by UFT methods, the datasets's format which have heterogeneous features can be combined to numerical features entirely. The solution for UFT methods is shown by equation(1) [13]. Based on equation(1), UFT methods can be formalized as shown by Algorithm 1, which also details equation(1) together.

$$\boldsymbol{\mu_i^*} = \left[(\boldsymbol{n}-\boldsymbol{i}) - \textstyle\sum_{k=1}^{i}(\boldsymbol{n}-\boldsymbol{k})\,\boldsymbol{p_k}\right]\sqrt{\left(\mathbf{1}-\textstyle\sum_i \boldsymbol{p_i^3}\right)/\textstyle\sum_{i\neq j}\boldsymbol{p_i}\,\boldsymbol{p_j}(\boldsymbol{i}-\boldsymbol{j})^2}, \quad (1)$$

where $\boldsymbol{\sigma_i^*} = \ \boldsymbol{p_i}\ \boldsymbol{i}\ \boldsymbol{\epsilon}\ \{\mathbf{1}, \ldots, \boldsymbol{n}\}$

Calculation of MI value between feature candidate and class label

MI is the amount of information that both variables share, and is defined as equation(2). Each feature f_i which is a member of *F* is calculated the value of MI (*I*) to class label *C*. By adopting equation(2), the value of MI for each feature f_i is obtained by using equation(3).

$$\boldsymbol{I(X;C) = \ H(C) - H(C|X)} \qquad (2)$$

Algorithm 1: UFT

Input: dataset *D*, which have heterogeneous feature f_j,*j* ϵ {1, ..., *m*}

Output: transformed dataset *D'* with pure numerical features

```
for j = 1 to m do
if feature f_j is non-numerical then
    $n = size(unique(f_j))$;
    $\{s_i | i = 1, ..., n\}$ is the set of non-numerical values in feature $f_j$
    $p_i$ is the probability of $s_i$
    for i = 1 to n do
$\mu_i = \left[(n-i) - \sum_{k=1}^{i}(n-k)p_k\right]\sqrt{(1-\sum_i p_i^3)/\sum_{i\neq j} p_i\, p_j (i-j)^2}$ ;
    $\sigma_i = p_i$ ;
        use Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i)$ to generate numerical data and substitute the values equal to $s_i$ in feature $f_j$
    end for
end if
end for
```

$$\boldsymbol{I(f_i;C) = H(C) - H(C|f_i)} \qquad (3)$$

I(f_i;*C*) = *H*(*C*) – *H*(*C*|f_i) where *H*(*C*) is defined as equation(4). *H*(*C*) is the entropy of class label *C*.

$$\boldsymbol{H(C) = -\textstyle\sum_{i=1}^{N} p(c_i)\log(p(c_i))} \qquad (4)$$

The value of *p*(c_i) probability function is obtained by using equation(5).

$$\boldsymbol{p(c_i) = \frac{number\ of\ instants\ with\ value\ c_i}{total\ number\ of\ instants\ (N)}} \qquad (5)$$

To fill the first subset, find *I*(f_i, *C*) which has a maximum value. Feature f_i is more relevant to the class label *C* than feature f_j in the context of the already selected subset *S* if it satisfies equation(6).

$$\mathbf{I(f_i, S; C) > I(f_j, S; C)} \qquad (6)$$

Calculate JMI value

Let *S* = {f_1, f_2, …, f_k}, JMI of f_i and each feature in *S* with *C* is calculated. The minimum value of this mutual information is selected based on the lowest amount of new information of feature *fi* that is added to subset. The feature that produces the maximum value is the feature that adds maximum information to that shared bet-ween *S* and *C*, it means that the feature is most relevant to the class label *C* in the context of the subset *S* according to equation(6).

The features are selected by JMIM according to equation(7), where JMI *I*(*fi*, *fs*; *C*) is defined as

Algorithm 2: Forward greedy search

```
1. (Initialisation) Set $F \leftarrow$ "initial set of $n$ features"; $S \leftarrow$ "empty set."
2. (Computation of the MI with the output class) For $\forall f_i \in F$ compute $I(C; f_i)$.
3. (Choice of the first feature) Find a feature $f_i$ that maximises $I(C; f_i)$; set $F \leftarrow F\backslash\{f_i\}$; set $S \leftarrow \{f_i\}$.
4. (Greedy selection) Repeat until $|S| = k$: (Selection of the next feature) Choose the feature $f_i = \arg max_{fi \subset F-S}(min_{fs \subset S}(I(f_i, f_s; C)))$; set $F \leftarrow F \setminus \{f_i\}$; set $S \leftarrow S \cup \{f_i\}$.
5. (Output) Output the set $S$ with the selected features.
```

equation(8).

$$\boldsymbol{f_{JMIM} = arg\, max_{f_i \in F-S}(min_{f_s \in S}(I(f_i, f_s; C)))} \quad (7)$$

$$\boldsymbol{I(f_i, f_c; C) = [-\textstyle\sum_{c \in C} p(c) log(p(c))] - \left[\textstyle\sum_{c \in C} \sum_{f_i \in F-S} \sum_{f_s \in S} log\left(\frac{p(fi\, fs;\, c/fs)}{p(fi/fs)\, p(c/fs)}\right)\right]} \quad (8)$$

Determine a feature subset

The method uses the following iterative forward greedy search algorithm to find the relevant feature subset of size *k* within the feature space (Algorithm 2).

Classification process

At this stage, classification process is conducted to determine the class of the object. In this study, the cclassification uses support vector machine (SVM) multiclass One-Against-One (OAO) with polynomial kernel. Polynomial kernel function (*K*) is shown by equation(9):

$$\boldsymbol{K(x, x_i) = [(x.x_i) + 1]^q} \quad (9)$$

where x_i is dimensional input (i = 1, 2, ..., l, l is the number of samples) belong to class 1 or another and q is power of polynomial kernel function.

Datasets

Datasets are used in this study from UCI Repository (table I). They are Acute Inflammations, Adult, Australian Credit Approval, German Credit Data, and Hepatitis. Data type of Acute Inflammations dataset is multivariate. The attribute types of this dataset are categorical and integer. This dataset contains 1 numerical feature and 5 non-numerical features. All of the non-numerical features only have two probability values, yes or no value. This dataset has two classes of data, they are yes for the inflammation of urinary bladder and no for not.

Data type of Adult dataset is multivariate. This dataset contains 14 features that composed by categorical and integer values. The attribute types of this dataset are categorical and number. Every feature has different number of values. This dataset has two data classes.

Australian Credit Approval dataset has multivariate data type. This dataset contains 14 features that composed by categorical, number, and real values. There are 6 numerical features and 8 categoical features. This dataset has two data classes. They are + (positive) class for approved credit and – (negative) class for rejected credit.

Data type of German Credit Data dataset is multivariate. This dataset contains 20 features that composed by categorical and number. There are 7 numerical features and 13 categorical features. This dataset has two classes of data, they are 1 as good credit consumer and 2 as bad credit consumer.

Data type of Hepatitis dataset is multivariate. The dataset contains 20 features that composed by categorical, number, and real values. There are 6 numerical features and 13 categorical features. This dataset has two classes of data, they are 1 for die and 2 for live.

3. Results and Analysis

To validate the results of proposed methods, five datasets from UCI Repository are used in the experiment (Table 1). In the datasets used, the type of non-numerical features is categorical data which is nominal and ordinal data type. The number of non-numerical features in each dataset is different (Table 2).

Scenario of testing is conducted by transforming non-numerical features using UFT methods and dummy variable. The transformation using dummy variable is conducted by changing the data to the numbers manually, for example feature of sex which has male and female data is changed by numeral 1 (for male) and 2 (for female).

It means Dummy-JMIM has lower complexity than UFT-JMIM but we do not know it is good for changing the categorical value manually or no. The transformed and the original features are combined entirely, then determine features subset by using JMIM methods, and classify them using SVM algorithm. The classification accuracy are measured for any number of selected feature subset and compared between UFT-JMIM methods and Dummy-JMIM methods.

Dummy variable is an defined variable whi-

TABLE 1
DESCRIPTION OF REAL-WORLD DATASETS

1	2	3	4	5	6
Acute Inflammations	120	2	1	5	6
Adult	1992	2	6	8	14
Australian Credit Approval	690	2	6	8	14
German Credit Data	1000	2	7	13	20
Hepatitis	80	2	6	13	19

Titles of column heads:
1: Datasets; 2: Instances; 3: Classes; 4: Numerical features
5: Non-numerical features; 6: Features

TABLE 2
NON-NUMERICAL FEATURES IN DATASETS

Datasets	Position of non-numerical features in datasets
Acute Inflammations	2, 3, 4, 5, 6
Adult	2, 4, 6, 7, 8, 9, 10, 14
Australian Credit Approval	2, 4, 5, 6, 8, 9, 11, 12
German Credit Data	1, 3, 4, 6, 7, 9, 10, 12, 14, 15, 17, 19, 20
Hepatitis	2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 19

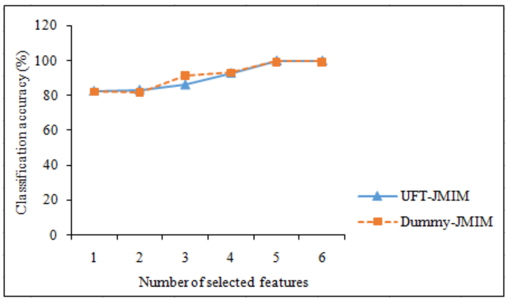

Figure. 4. The classification accuracy that is achieved by the Acute Inflammations dataset

ch is created to represent an attribute with two or more different categories or levels. We use dummy variable as another way for the transformation of features by using defined variable. Dummy variable is used as a reference to ensure that the results of transformation from non-numerical into numerical features by using UFT methods does not have significant difference to the results of transformation of features using defined variable. So that, indicating that the original feature information is not lost.

Figures 2-6 show the classification accuracy of the five datasets. The classification accuracy is computed for the whole size of the selected subset (from 1 feature up to 20 features). Thus, all features of each dataset in this experiment was selected for each testing of *k* value (number of selected features). As shown in Figure 2, it illustrates the experiment with the acute inflammations dataset. UFT-JMIM achieves the highest average accuracy (100%) with 5 and 6 selected features, which is higher than the accuracy of Dummy-JMIM with 5 features (99.5%) and 6 features (99.4%).

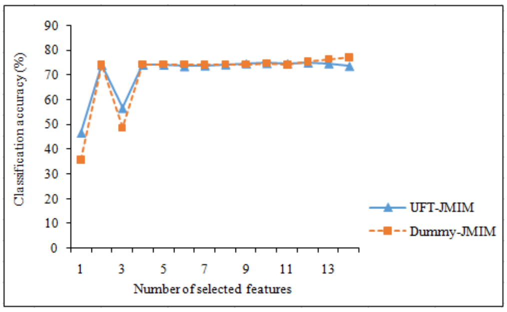

Figure. 3. The classification accuracy that is achieved by the Adult dataset

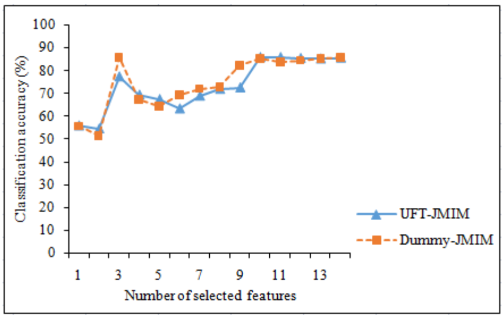

Figure. 2. The classification accuracy that is achieved by the Australian Credit Approval dataset

In Figure. 3 which illustates the accuracy of the adult dataset, UFT-JMIM cannot achieve the highest classification accuracy. It can only achieve the 74.9% with 12 selected features. Meanwhile, Dummy-JMIM can achieve classification accuracy 77.1% with 14 features. Figure 4 shows the results for australian credit approval dataset. UFT-JMIM can achieve the highest classification result (85.83%) with 11 selected features. Dummy-JMIM can achieve the closest classification accuracy (85.80%) with 14 features.

The classification accuracy of UFT-JMIM for german credit data dataset is shown by Figure 5. It achieves 72.3% (15 selected features). Whereas, the classification accuracy produced by Dummy-JMIM can only achieve 70.6% as the best result with 1 selected feature. Figure 6 shows the UFT-JMIM performance for the hepatitis dataset which achieves the highest classification accuracy (89.3%) with 10 selected features. Meanwhile, Dummy-JMIM can only achieve 88.2% with 10 selected features.

UFT-JMIM can get better classification accuracy because of MI that used in UFT and JMIM methods. In UFT methods, MI preserves the information of data when transformation of features is conducted from non-numerical to numerical features. So when the data is transformed, MI minimizes information loss. For this case, the MI value between the transformed and the original features must be the same as the entropy of the ori-

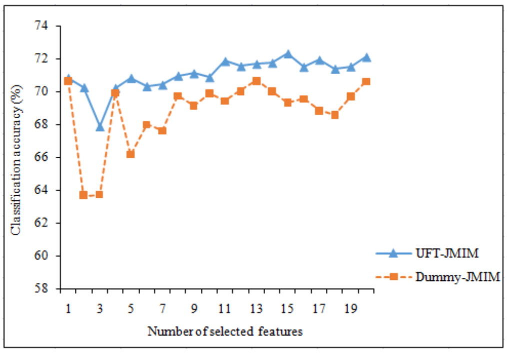

Figure. 6. The classification accuracy that is achieved by the German Credit Data dataset

ginal features to preserve the original feature information.

In addition, MI applied in JMIM methods is used to measure of relevant and redundant features when select feature subset. It studies relevancy and redundancy, and takes into consideration the class label when calculating MI. In this methods, the candidate feature that maximises the cumulative summation of joint mutual information with features of the selected subset is chosen and added to the subset. JMIM methods employs joint mutual information and the 'maximum of the minimum' approach, which should choose the most relevant features. The features are selected by JMIM according to criterion as equation(7). In JMIM methods, the iterative forward greedy search algorithm is used to find the best combination of *k* features within subset. It causes the performance of finding to feature subset to be suboptimal because of high computation.

4. Conclusion

Feature selection based on MI using trans-formed features can reduce the redundancy of the selected feature subset, so that it can improve the accuracy of classification. The average classification accuracy for all experiments in this study that can be achieved by UFT-JMIM methods is about 84.47% and Dummy-JMIM methods is about 84.24%. This result shows that UFT-JMIM methods can minimize information loss between transformed and original features, and select feature subset to avoid redundant and irrelevant features.

For future work, further improvement can be made by studying to determine the best size *k* to find the relevant feature subset from heterogeneous features automatically in which it may make computation to be low.

References

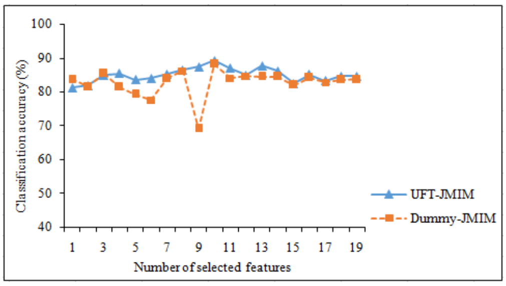

Figure. 5. The classification accuracy that is achieved by the Hepatitis dataset

[1] Bajwa, I., Naweed, M., Asif, M., & Hyder, S., "Feature based image classification by using principal component analysis," *ICGST International Journal on Graphics Vision and Image Processing*, vol. 9, pp. 11–17. 2009.

[2] Turk, M., & Pentland, A., "Eigenfaces for recognition," *Journal of Cognitive Neuro-science*, vol. 3, pp. 72–86. 1991.

[3] Tang, E. K., Suganthana, P. N., Yao, X., & Qina, A. K., "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognition*, vol. 38, pp. 485–493. 2005.

[4] Yu, H., & Yang, J., "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070. 2001.

[5] Bennasar, M., Hicks, Y., & Setchi, Rossitza, "Feature selection using Joint Mutual Information Maximisation," *Expert Systems With Applications*, vol. 42, pp. 8520-8532. 2015.

[6] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L.A., *Feature extraction foundations and applications*, Springer Studies infuzzi--ness and soft computing, New York/Berlin, Heidelberg, 2006.

[7] Saeys, Y., Inza, I., & Larrañaga, P., "A review of feature selection techniques in bioinformatics," *Bioinformatics Advance Access*, vol. 23, pp. 2507-2517. 2007.

[8] Battiti, R., "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550. 1994.

[9] Peng, H., Long, F., & Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238. 2005.

[10] Estévez, P.A., Tesmer, M., Perez, A., & Zurada, J. M., "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, pp. 189–201. 2009.

[11] Hoque, N., Bhattacharyya, D. K., & Kalita, J. K., "MIFS-ND: a mutual information based feature selection method," *Expert Systems with Applications,* vol. 41, pp. 6371–6385. 2014.

[12] Yang, H., & Moody, J., "Feature selection based on joint mutual information" *In Proceedings of international ICSC sympo-sium on advances in intelligent data analysis*, pp. 22–25, 1999.

[13] Wei, M., Chow, T. W.S., & Chan, R. H.M., "Heterogeneous feature subset selection using mutual information-based feature transformation," *Neurocomputing*, vol. 168, pp. 706-718. 2015.