Feature Selection via Correlation Coefficient Clustering

Hui-Huang Hsu
Department of Computer Science and Information Engineering, Tamkang University, Taipei, 25137, Taiwan
Email: h_hsu@mail.tku.edu.tw

Cheng-Wei Hsieh
Department of Computer Science and Information Engineering, Tamkang University, Taipei, 25137, Taiwan
Email: 892190108@s92.tku.edu.tw

***Abstract*—Feature selection is a fundamental problem in machine learning and data mining. How to choose the most problem-related features from a set of collected features is essential. In this paper, a novel method using correlation coefficient clustering in removing similar/redundant features is proposed. The collected features are grouped into clusters by measuring their correlation coefficient values. The most class-dependent feature in each cluster is retained while others in the same cluster are removed. Thus, the most class-related and mutually unrelated features are identified. The proposed method was applied to two datasets: the disordered protein dataset and the Arrhythmia (ARR) dataset. The experimental results show that the method is superior to other feature selection methods in speed and/or accuracy. Detail discussions are given in the paper.**

***Index Terms*—Feature Selection, Clustering, Correlation Coefficient, Support Vector Machines (SVMs), Machine Learning, Classification**

I. INTRODUCTION

Feature selection aims to select the most problem-related features and to remove unnecessary features [1]. The unnecessary features include both noisy and redundant features. We can say that if a feature cannot help improve the classification accuracy, the feature is useless and unnecessary. The noisy feature is especially meant to harm the classification results. If the class classification result is improved by removing some features, we can say that these features could be noisy features. But one important question is how to find these noisy features? The wrapper mode feature selection model could be helpful [2]. However, it is usually very time consuming, because it combines some learning machines which are the core of selecting features [3][4]. Features which lower the overall accuracy by the learning machine will be removed from the original feature set. The procedure would be progressively repeated until the classification accuracy cannot be further improved. This procedure needs complicated computation and always takes a lot of time.

In this paper we focus on reducing repeated or redundant features. The targeting features may not be exactly the same, but they are closely related. Similar features inputted to the classifier not only increase the computation time, but also decrease its classification capability. There are several measures which are helpful in finding the redundant features. For example, mutual information, correlation coefficient, and chi-square can be used to find the dependency between two features. However, for a large amount of features, this pairwise dependency information is not enough for us to find the features which are close to each other in groups. Hence, clustering analysis is applied here. It is a very useful technique to divide a feature set into subsets within which features are closely related to each other. If we can separate the collected features into such groups, we need to keep only one feature in each group because they are almost the same. Therefore we can greatly reduce the number of features by removing those redundant features.

Clustering analysis usually uses Euclidean distance as the similarity measurement. But measurements based on the information theory could be more helpful in finding dependency between two variables than simply measuring the distance in space. In this research, the correlation coefficient instead of the Euclidean distance is used for clustering analysis. The correlation coefficient of two random variables is a quantity that measures the mutual dependency of the two variables. Hence, when two features are mutually dependent, it means the occurrence and variation of the two features must be almost the same. For a classification problem, we need to keep only one of them since they share almost the same characteristics.

For hundreds or even thousands of collected features, there must be features that are very similar to each other, and we can take these features as the same kind of features. We certainly do not need to use all features of the same kind for classification. After clustering analysis identifies all different kinds of features, we can remove a great number of redundant features. The classification performance in both the computational speed and the classification accuracy can be improved with the removal of these redundant features. A novel feature selection algorithm based on the above-mentioned correlation coefficient clustering is proposed in this paper. Support vector machines (SVMs) [5] are used as the classifier for testing the feature selection results on two datasets:

doi:10.4304/jsw.5.12.1371-1377

disordered protein data and Arrhythmia (ARR) data. Details are given in the subsequent sections.

The rest of this paper is organized as follows. Section II introduces related work. Section III presents the proposed clustering feature selection mechanism. Section IV describes the SVM learning model and the datasets. Section V shows experimental results and discussions. Finally, Section VI draws a brief conclusion.

II. Related Work

Feature selection methods have been applied to classification problems in order to select a reduced feature set that makes the classifier faster and more accurate. Roughly speaking, the feature selection model contains two different modes: filters and wrappers [2]. The filters measure the information of features [6][7] (e.g., information gain) to decide the feature selection result. This kind of model works fast, but the classification result is not always satisfied. Because the filters contain no error rate controlling technique, the result of filters is not always stable. On the other hand, the wrappers combine a learning model in it. The wrappers perform the feature selection through two main steps: feature searching and classification error rate measurement. The feature searching procedure selects features from the original feature set and input them into the next classification procedure to test their prediction error rate. The wrappers work slowly because both the two main steps are very time-consuming. Moreover, complex calculation makes it difficult to perform the wrappers on applications with a large number of features.

In our previous research, we combined the filters and the wrappers to solve the applications with a large number of features [8]. At first, we use the fast filter models with two information measurement: information gain and F-score. These two models can filter out a lot of features not that related to the problem. As mentioned above, the filter might not provide a satisfied classification result. Hence, we perform the wrapper-mode feature selection to improve the classifier's prediction result. The hybrid mechanism was applied to the protein disordered region prediction problem which is to find out the unstructured regions of proteins. The learning model used in it was the support vector machine. In the experimental results, 350 features were selected from the original 440 features and the prediction accuracy was 82.72%.

One way to solve the problem of redundant or repeated features is to use some kind of feature dependency measurements, such as mutual information (MI), correlation coefficient, or chi-square. A mutual information feature selection mechanism was proposed by Huang et al. [9]. They used a filter approach to perform the feature selection. In their point of view, there are two types of input features perceived as being unnecessary. They are features completely irrelevant to the output classes and features redundant given other input features. By using the mutual information performed on class-related and feature-related features, feature selection can be done. The concept is from the information theory which analyzes the relationship between features and classes to remove the redundant features and the most irrelevant features to the class.

Another feature dependency measurement feature selection was proposed by Peng et al. [10]. They also used mutual information to perform feature selection. Their original feature selection concept is based on features max dependency (MaxDep) [11] which measures the feature sets' statistical dependency with the target class. MaxDep selects m features that jointly have the largest dependency on the target class. The final selected features have the maximal dependency values that are calculated from some similarity measurements, for example, correlation coefficient or mutual information. However, the estimation of MaxDep is very hard due to its multivariate dependency measurement which is retrieved from a high dimensional space. Both feature searching and information measuring are quite time-consuming. In order to improve MaxDep, Peng et al. designed a two-stage feature selection algorithm by combining the minimal-redundancy-maximal-relevance criterion (mRMR) and other more sophisticated feature selectors. It calculates the features with the maximal class-related value while this feature is in the minimal redundancy with all the already selected features. It then performs optimal first-order incremental selection to improve the classification result. By using some wrapper kind of feature selection model (e.g., forward/backward floating search), they get the final compact feature set with the highest classification accuracy. The results confirm that mRMR leads to promising improvement on feature selection and classification accuracy.

For the feature dependency measurement techniques, the correlation coefficient also plays an important role though it has not been used as often as mutual information. From the definition, the correlation coefficient provides a quantitative measurement that represents the strength of a linear relationship between two sequences of observations. Hence, for most variables relationship tests, calculating correlation coefficients would be the first step to determine if they are linearly dependent. On the other hand, mutual information is based on the knowledge measurement, which handles the test of how much knowledge one can gain of a certain variable by knowing the value of another variable. Mutual information helps reduce the range of the probability density function for a random variable x if the variable y is known. Therefore, if we only want to test the dependency between two variables instead of testing the knowledge gain, it is preferable to use the correlation coefficient. In the next section, we introduce the correlation coefficient based feature selection model which can find out redundant features by testing pairwise feature dependency.

III. Correlation Coefficient Clustering for Feature Selection

To find related feature groups is not an easy task. The pairwise similarity measurements of the whole feature set are hard to be realized due to a large amount of huge

calculations. Besides, the result of pairwise measurements cannot be used to identify multiple similar features. Thus we propose to use clustering analysis to group the most related features together. This could divide the feature set into groups of multiple features.

The Euclidean distance is the most used similarity measurement in clustering analysis. However, it does not fit our feature selection goal. Therefore, we replace the distance measurement with the correlation coefficient in clustering. Next, feature selection within feature clusters is also an important problem. This is also an important procedure of feature selection. One representative feature needs to be picked from each feature cluster. In previous researches, little attention was paid to this problem. The researchers thought that since the features in the same cluster are almost the same, any of them can be chosen and the classification results would be about the same. But there exists difference among those similar features. Here we propose to choose the feature most related to the class in each feature cluster. The feature that has the highest correlation coefficient value with the class label is picked.

The following subsections introduce the clustering mechanism, the correlation coefficient, and the proposed correlation coefficient clustering algorithm for feature selection.

A. Clustering

Clustering is one of the most widely used techniques for exploratory data analysis. It also can be considered as the most important unsupervised learning problem. Practically, clustering analysis finds a structure in a collection of unlabeled data. Hence, it separates the original dataset into smaller datasets called *clusters*. Data in each cluster are close to each other. Fig.1 demonstrates such separation of data.

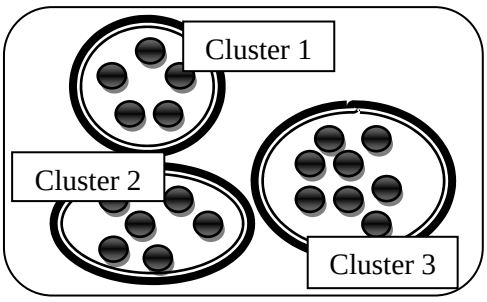

Figure 1. Separation of data via clustering.

Clustering algorithms can be classified as hierarchical clustering, overlapping clustering, exclusive clustering, and probabilistic clustering [12]. In our research, we only consider the exclusive clustering, and that means each node in the Fig.1 can only belong to one cluster. There are also many clustering algorithms. Among them, K-means is the classical one. For K-means clustering, it works on separating n observations into k clusters, and each observation belongs to the nearest mean's cluster. Usually the Euclidean distance is used as the distance metric to calculate the observations' relationship. K-means clustering works as the following steps.

1. Randomly select k nodes as the *means* from n observations, where $k \leqq n$.
2. Calculate the Euclidean distance from each node to all the means, and the (n-k) observations belong to their respective nearest mean.
3. Re-calculate the means of all clusters $m_1, m_2, \ldots, m_k$.
4. Repeat Steps 2 and 3 until the content of each cluster is fixed.

Finally, each cluster could represent a different collection from the other clusters. By using this kind of clustering models, the observations could be easily separated according to the Euclidean distance measurement. This is much better than measuring the distance between each pairs for all the observations considering the time complexity. However, the Euclidean distance can only measure the space distance between observations. The observations' dependency cannot be revealed. Hence, in this paper, we apply the correlation coefficient in clustering to measure the dependency of all observations.

B. Correlation Coefficient

In statistics, the correlation coefficient indicates the strength and direction of a relationship between two random variables. The commonest use refers to a linear relationship. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. Equation (1) shows the calculation of the correlation coefficient between two variables x and y. There are totally n observations.

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2}} \tag{1}$$

Two variables have strong dependency when their correlation coefficient value is close to 1 or -1. When the value is 0, it means that the two variables are not related at all. In our research, strong dependency is what we are looking for, no matter it is positive or negative. Therefore, in the measurement procedure, the absolute value of the correlation coefficient $| r |$ is used.

C. Correlation Coefficient Clustering Algorithm

In this study, we combine the correlation coefficient with clustering analysis for feature selection. Instead of using the Euclidean distance, we choose the correlation coefficient as the similarity measurement as discussed in the previous subsection. Moreover, clustering analysis can separate the whole feature set into different groups. Closely related features can be put together after the first clustering steps. The features are divided into different kinds of groups according to their dependency. And each kind of groups can represent a part of the feature space.

For the final goal of feature selection, we must choose the most relevant and non-redundant features from the original feature set to reduce the number of features. In this approach, only one feature is needed from each kind/cluster of features. The reason is that features in the

same cluster are very close to each other and we do not need to use more than two features of the same kind to perform the classification task. Fig. 2 shows the concept of the proposed feature selection model. In the clustering procedure, we use the correlation coefficient as the similarity measurement to check the dependency among features.

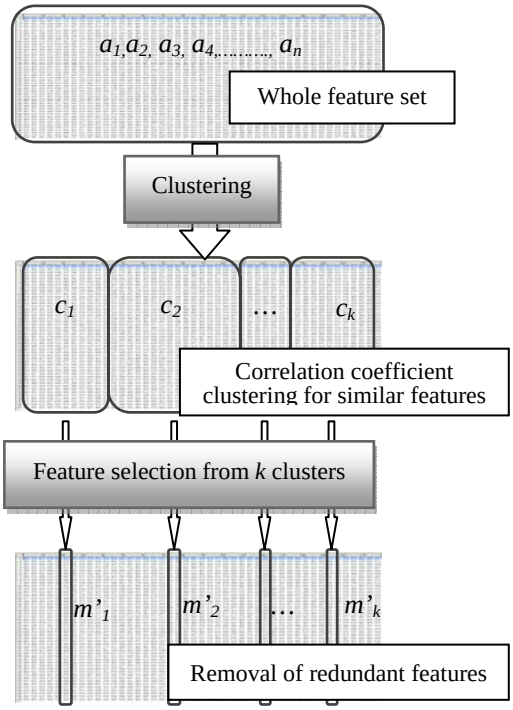

Figure 2. The process of correlation coefficient clustering feature selection. The remained features are the result of feature selection.

A problem comes up here regarding how to pick the representative feature for each feature cluster. That is, which feature in a cluster should we keep? We propose to pick the most class-dependent feature in each cluster as the representative one. The correlation coefficient can also be used to decide the class-feature dependency. The most class-dependent features from all clusters can certainly help improve the overall classification accuracy. The pseudocode of the proposed correlation coefficient clustering feature selection algorithm is as follows.

```
Randomly select k nodes m(m1,…, mk) from n observations a(a1,…,an);
  WHILE originally selected k nodes m(m1,…, mk)  != new selected k'
          nodes m'(m1',…, mk')
    FOR i = 1 to n (observations)
      FOR j = 1 to k (nodes)
          rj = Correlation_Coefficient (ai, mj);
          IF rj    MAX(r1, r2,…, rk)
            ai belongs to mj's cluster;
          END IF
      END FOR
    END FOR
    FOR p = 1 to k (clusters C1,…,Ck)
      FOR q = 1 to Cp's length t(cluster p's contents s1,…,st)
          rq = Correlation_Coefficient (sq, Class labels);
          IF rq    MAX(r1, r2,…, rp)
            mp'= sq;
          END IF
       END FOR
    END FOR
  END WHILE
RETURN m'(m1',…, mk') ;
```

Next, we make a brief comparison of the proposed method with mRMR. First, mRMR only choose the most informational features, i.e., the most class-related features. As we know the "the *m* best features are not the best *m* features" [13], the result by mRMR might ignore features which are not so closely related to the class label, but can complement other features to improve the classification result. In the proposed method, no such features would be missed. Secondly, the Min-Redundancy step of mRMR only randomly keeps one of the Max-Relevance features. On the other hand, the proposed method retains the most class-related feature in each feature cluster by calculating the correlation coefficients between the features and the class. Other features in the same cluster are then removed.

IV. Learning Model and Datasets

A machine learning method is needed when we apply the proposed feature selection in classification problems. The support vector machine (SVM) was chosen for the experiments in this research due to its advantages in the use of kernels for nonlinear problems and the optimization of the separating margins. Furthermore, it can avoid the local minima problems during the training process. In this section, the datasets used for the experiments in this research are also introduced.

A. Support Vector Machine

The SVM is based on the SV (support vector) learning. That means the SVM does not always compare the prediction target with all the existing training nodes. In contrast, the SVM selects a group of nodes as its SVs, and uses these SVs to judge the label of the classification target. In the testing stage, the SVM model uses the SVs to do the classification. These SVs locate near the hyperplanes that cause the maximum margin of class separation. Fig. 3 demonstrates the maximum margin between two classes which are separated by the hyperplane in the SVM model. H_1 and H_2 are the boundaries. And the nodes which are located near these two lines are support vectors.

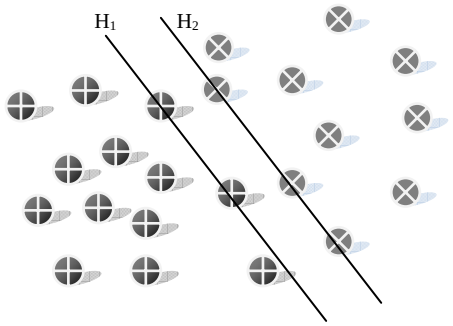

Figure 3. The SVM could find out the maximum margin and use the SVs to predict the prediction targets. Boundaries H1 and H2 are located on these SVs.

B. Datasets

Protein disordered region prediction is the first problem tried in this research. In proteomics, a protein's function is always strongly related to its structure. While some parts of a protein have a fixed definite structure, such as α-helix, β-sheet, or coil, other parts are not associated with well-defined conformations. Previously, these so-called disordered regions were not thought to have a specific function of their own. But, recent studies suggest that some disordered regions may have important signaling or regulatory functions. In addition, some critical diseases are strongly related to these disordered regions. Thus, protein disordered region prediction is an important problem. However, the most relevant features in this problem are yet to be determined [14][15].

Our ordered and disordered sequences were collected from the PDB [16] and DisProt [17] databases. The proteins in DisProt are all with disordered regions. The protein sequences collected from PDB contain mostly ordered regions. Those data selected from DisProt are taken as positive training data, and the negative training data are derived from PDB_Select_25 [18] which is a non-redundant dataset of the Protein Data Bank (PDB). Finally, 119 protein sequences with 440 features were collected and there are totally 21676 residues. The 440 features were determined from related researches [8].

In order to compare the proposed method with MaxDep and mRMR, the Arrhythmia (ARR) dataset from UCI machine learning archive [19] was also used. The aim of this dataset is to distinguish between the absence and presence of cardiac arrhythmia and to classify a datum into one of the 16 classes. However, we can only consider two states: normal and abnormal. Class 1 refers to *normal*, Classes 2 to 15 refer to different *abnormal* classes of arrhythmia, and Class 16 refers to the other unclassified ones. In this dataset, there are totally 452 instances with 279 features. Among the features, 206 are linear values and the rest are nominal.

V. Experimental Results

A software tool has been implemented for the proposed feature selection method (Fig. 4). C#.NET in MS Visual Studio was used to develop the tool. The user can determine the number of clusters, the similarity measurement, and the clustering method in our tool.

As for the determination of the number of clusters in the experiments, several methods have been tried, namely, gap statistic [20], Calinski-Harabasz index [21], Krzanowski-Lai index [22], and Hartigan statistic [23]. Most of them compare the values of between-cluster sums of squares and the values of within-cluster sums of squares to detect the distribution of data. Following their distribution, the number of clusters can be estimated. There are two main problems. First, these methods can only give estimates and sometimes perform not so precisely. Secondly, in our experiment, the number of clusters is also the final number of remained features. According to the past researches, with only main class-related features the classifier might not perform well. Sometimes it is necessary to include some additional features to improve the classifier's discrimination ability. Therefore, in our experiments, although we had the estimated number of clusters from these models, we still tried several different numbers of clusters.

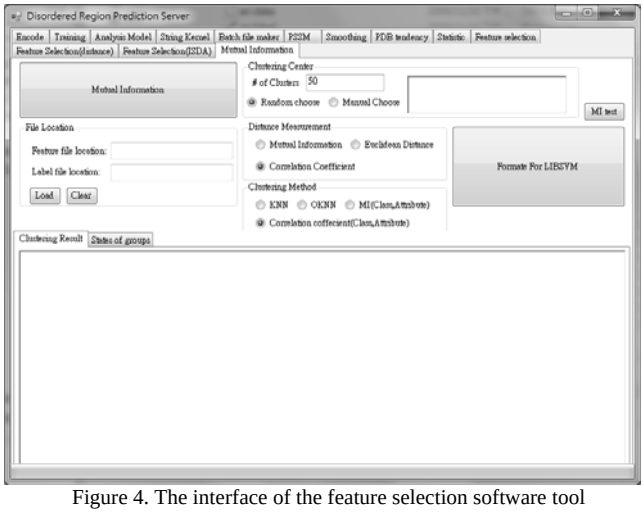

Figure 4. The interface of the feature selection software tool

For the SVM learning machine in this experiment, we use the RBF kernel. The experimental results of protein disordered region prediction with the proposed method are listed in TABLE 1. There are totally 440 features in the original dataset. The best result via five-fold cross-validation is 86.30% with only 200 features. It is much better than the result produced by our previous work with a hybrid feature selection model [8]. The best result in [8] was 82.72% with 350 features. The number of features is further reduced by 34% ((350-200)/440) and the classification accuracy is raised by 3.58%. This demonstrates the usefulness of the proposed feature selection method.

TABLE 1.
FIVE-FOLD CROSS-VALIDATION RESULTS ON DISORDERED PROTEIN DATA

Feature number	Accuracy (5-fold cross-validation)
50	82.28%
100	84.00%
150	85.67%
200	86.30%

Next, we compare the proposed method with mRMR and MaxDep [10] on the ARR dataset. Fig. 5 shows that the proposed method is better than MaxDep and comparable to mRMR in classification accuracy. The number of selected features ranges from 5 to 55 (from the original 279 features). The proposed method provides a better and more stable result than MaxDep. In the procedure of feature searching, MaxDep has to search through the whole feature set with different combinations. This procedure also takes time.

The proposed method did not perform better than mRMR. The reason is that mRMR incorporates the wrapper mode in the second stage of its feature selection

procedure. The wrapper mode works as a post modification step which can further improve the classification accuracy by repeatedly using a learning machine. This repeated process is very time-consuming. On the other hand, our method only uses clustering analysis once. It is more like a filter mode feature selection procedure that does not require very complex calculations.

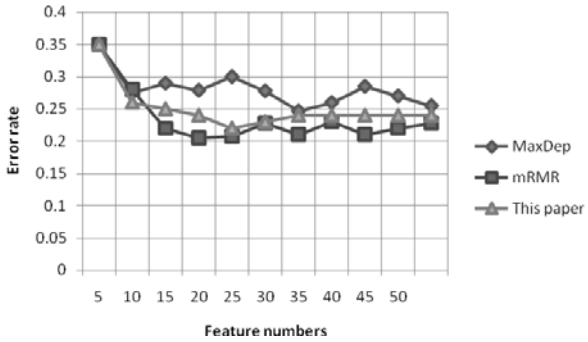

Figure 5. Ten-fold cross-validation accuracy comparison among MaxDep, mRMR, and correlation coefficient clustering feature selection on Arrhythmia data (learning machine: SVM)

From the experimental results, we can observe that the number of features can be greatly reduced by the proposed method on both datasets. The advantage of the proposed method is that it can execute much faster than the wrapper-mode feature selection methods while maintaining comparable classification accuracy. Clustering analysis is very helpful in finding maximal dependency among features. Each cluster can represent a different kind of features.

VI. Conclusion

In this paper, a novel feature selection method is proposed. The key characteristic of the method is to apply clustering analysis in grouping the collected features. Only one representative feature is needed from each feature group. This can greatly reduce the total number of features. In the method, the correlation coefficient is used to find similar features with maximum dependency. It is also used to identify the most class-dependent feature as the representative feature in each feature cluster.

Filter-mode feature selection methods only focus on identifying the most class-related features without considering redundancy among these features. Also, some removed features are actually helpful to the overall classification performance, but are viewed as not so class-related and removed just because their measures are low. On the other hand, feature selection methods involved with the wrapper mode require a lot of computations. The proposed method is advantageous to both filter-mode and wrapper-mode methods.

This method is yet to consider the removal of noisy features which can be harmful to the overall performance. One simple way to identify possible noisy data is to look for the representative features which have a *low* correlation coefficient value with the class. A representative feature with a near zero correlation coefficient value should definitely be removed. But experiments are needed to carefully examine the threshold setting. This is one future direction of this research.

References

[1] C. Deisy, B. Subbulakshmi, S. Baskar, and N. Ramaraj, "Efficient Dimensionality Reduction Approaches for Feature Selection," *International Conference on Computational Intelligence and Multimedia Applications*, vol. 2, pp. 121- 127, 2007. [doi: 10.1109/ICCIMA.2007.288]

[2] R. Kohavi, and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, vol. 97, pp. 273-324, 1997. [doi: 10.1016/S0004-3702(97)00043-X]

[3] J. R. Quinlan, *Discovering Rules from Large Collections of Examples: A Case Study*, In Michie, D. ed., Expert Systems in the Microelectronic Age, Scotland: Edinburgh University Press, Edinburgh, 1979, pp. 168-201.

[4] Y. Liu, Y. F. Yin, J. J. Gao, and C. G. Tan, "Wrapper Feature Selection Optimized SVM Model for Demand Forecasting," *The International Conference on Young Computer Scientists*, pp. 953-958, 2008. [doi: 10.1109/ICYCS.2008.151]

[5] LIBSVM - A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed Nov 3, 2009)

[6] A. Al-Ani, "A dependency-based search strategy for feature selection," *Expert Systems with Applications: An International Journal*, vol.36, pp. 12392-12398, 2009.

[7] B. Bonev, F. Escolano and M. Angel-Cazorla, "A Novel Information Theory Method for Filter Feature Selection," *MICAI 2007: Advances in Artificial Intelligence*, Springer Berlin / Heidelberg, pp. 431-440, 2007.

[8] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "A Hybrid Feature Selection Mechanism," *in Proc. Eighth International Conference on Intelligent Systems Design and Applications (ISDA 2008)*, vol. 2, pp. 271-276, Kaohsiung, Taiwan, Nov. 26-28, 2008. [doi: 10.1109/ISDA.2008.280]

[9] J. J. Huang, Y. Z. Cai, and X. M. Xu, "A Filter Approach to Feature Selection Based on Mutual Information," *Cognitive Informatics*, vol. 1, pp. 84 -89, 2006. [doi: 10.1109/COGINF.2006.365681]

[10] H. C. Peng, F. H. Long and C. Ding, "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, no. 8, pp. 1226-1238, 2005. [doi: 10.1109/TPAMI.2005.159]

[11] C. Ding and H. C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Proc. Second IEEE Computational Systems Bioinformatics Conf.*, pp. 523-528, 2003. [doi: 10.1109/CSB.2003.1227396]

[12] M. Matteucci, "A Tutorial on Clustering Algorithms", http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/ (last accessed Feb 3, 2010)

[13] T. M. Cover, "The Best Two Independent Measurements Are Not the Two Best," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 4, pp. 116-117, 1974.

[14] C. Bracken, L. M. Iakoucheva, P.R. Romero, and A.K. Dunker, "Combining prediction, computation and

experiment for the characterization of protein disorder," *Curr. Opin. Struct. Biol*, vol. 14, pp. 570-576, 2004.
[15] K. Peng, P. Radovojac, S. Vucetic, A. K. Dunker and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinformatics*, vol. 7, pp. 208, 2006. [doi: 10.1186/1471-2105-7-208]
[16] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig et al., "The Protein Data Bank," *Nucleic Acids Resource*, vol.28, pp. 235-242, 2000. [doi:10.1107/S0907444902003451]
[17] S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva et al., "DisProt: A Database of Protein Disorder," *Bioinformatics*, vol 21, pp. 137-140, 2005. [doi: 10.1093/bioinformatics/bth476]
[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403-410, 1990. [doi:10.1006/jmbi.1990.9999]
[19] UCI machine learning repository, http://www.ics.uci.edu/mlearn/mlsummary.html (last accessed Feb 3, 2010)
[20] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistics," *Journal of the Royal Statistical Society*, Series B 63, pp. 411-423, 2001.
[21] R. B. Calinski, and J. A. Harabasz, "A denrite method for cluster analysis," *Communications in Statistics*, vol. 3, pp. 1-27, 1974.
[22] L. Kaufman, and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.
[23] J. A. Hartigan, *Clustering Algorithms*. Wiley, 1975.

Hui-Huang Hsu is an Associate Professor in the Department of Computer Science and Information Engineering at Tamkang University, Taipei, Taiwan. He received both his PhD and MS Degrees from the Department of Electrical and Computer Engineering at the University of Florida, USA, in 1994 and 1991, respectively.

He has published over 80 referred papers and book chapters, as well as participated in many international academic activities. His current research interests are in the areas of machine learning, data mining, bio-medical informatics, ambient intelligence, and multimedia processing. He is a senior member of the IEEE.

Cheng-Wei Hsieh received his master's degree in Computer Science and Information Engineering at National Central University. His MS degree is from the Department of Computer Science & Information Engineering at Tamkang University, Taipei, Taiwan.

He is a PhD candidate in the Department of Computer Science & Information Engineering at Tamkang University, Taipei, Taiwan. His major research interests include applications in bioinformatics, machine learning.