# Discriminative Frequent Pattern Analysis for Effective Classification

Hong Cheng, Xifeng Yan, Jiawei Han, Chih-Wei Hsu

Presented by:
Abhishek Srivastava

CMPUT 695 Class Presentation
Department of Computing Science
University of Alberta

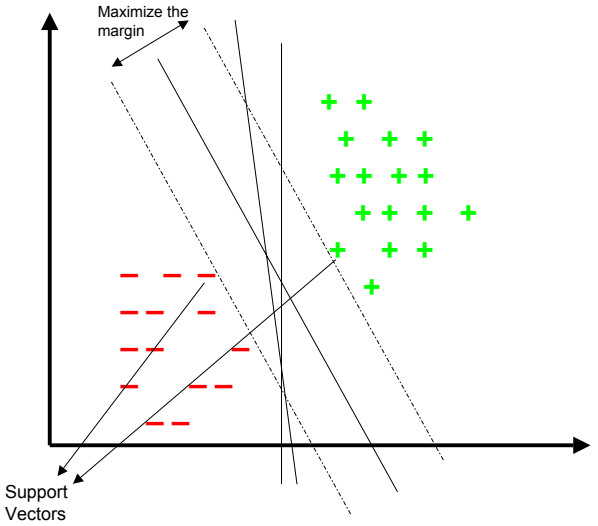
## Introduction

Cheng et al. in this paper put forward a new method for feature space conversion of the input to classifiers.

Their method, which mainly makes use of frequent patterns is typically suitable for Support vector machine (SVM) although they demonstrate its effectiveness on other classifiers as well.

The paper is a description of this method along with adequate justification for every step.
This is followed by extensive experimental results to substantiate their claims.

## Support Vector Machine (SVM)




Linearly separable data

## SVM

The hyper-plane of an SVM may be represented by : $(W.X)-b=0$

The margins may be represented by : $(W.X)-b=1$

$(W.X)-b=-1$

A data point $x_i$ in the positive class is given by : $(W.X_i)-b\geq 1$

A point in the negative class : $(W.X_i)-b\leq -1$

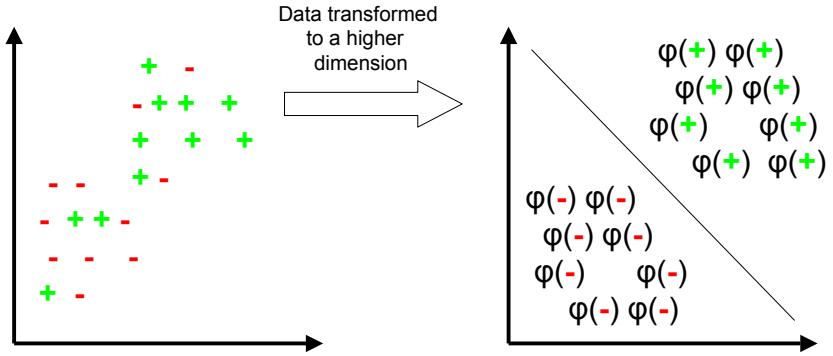
$$\Rightarrow \quad c_i\{(W.X_i)-b\}\geq 1$$

This expression in the dual form may be written as :

$$d(X^T)=\sum_{i=1}^{l} y_i\alpha_i(X_i.X^T)+b_o$$

SVM

Linearly inseparable data




SVM

A simple example demonstrating transformation of data to a slightly higher dimension :

$X = (x_1, x_2, x_3)$

⇩

$Z : \varphi_1(X) = x_1$
$\varphi_2(X) = x_2$
$\varphi_3(X) = x_3$
$\varphi_4(X) = x_1x_2$
$\varphi_5(X) = (x_3)^2$
$\varphi_6(X) = x_1x_2x_3$

SVM

A test-point $\varphi^T(X)$ in this higher dimension may be classified based on the following :

$$d(\varphi^T(X)) = \sum_{i=1}^{l} y_i \alpha_i (\varphi_i(X).\varphi^T(X)) + b_o$$

⇩

$$\varphi_i(X).\varphi^T(X)$$

⇩

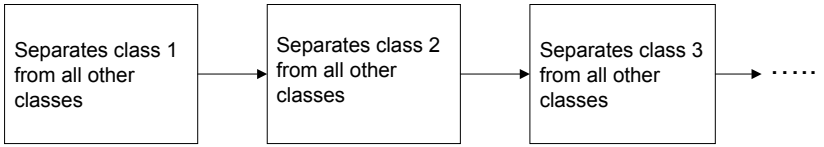
$$K(X_i, X_j)$$

(Kernel Function)

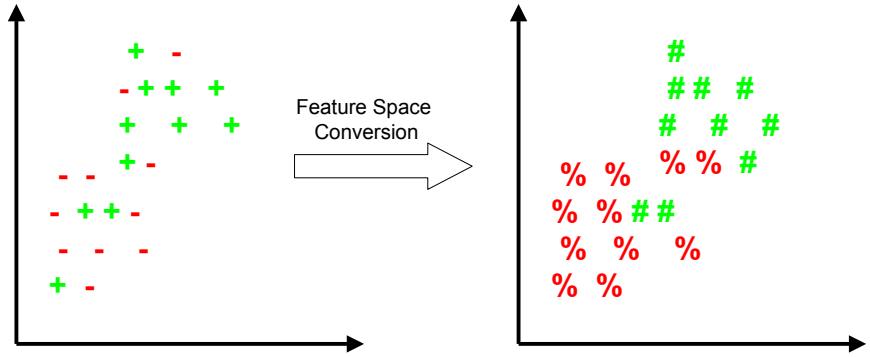
SVM

Typically SVM is used for data with only two classes.

It may however be used for multiple classes in the following manner :



## Feature Space Conversion




Feature space conversion somewhat improves the “separability” of the data so that the classifier has a relatively easier task.
This effectively improves the accuracy of the classifier.

## Feature Space Conversion

A simple example demonstrating feature space conversion :

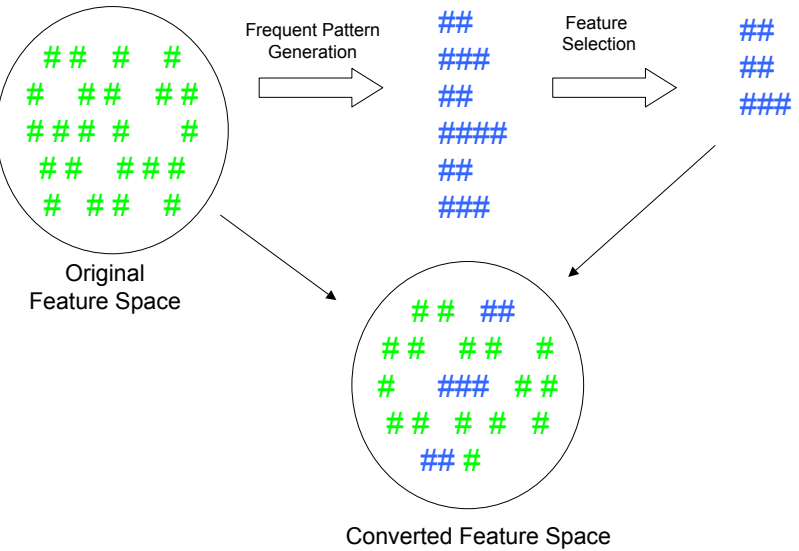
$$f(m_1, m_2, r) = C\frac{m_1 m_2}{r^2}$$

$$(m_1, m_2, r) \longrightarrow (x, y, z) \longrightarrow (\ln m_1, \ln m_2, \ln r)$$

$$\ln f(m_1, m_2, r) = \ln C + \ln m_1 + \ln m_2 + \ln r$$

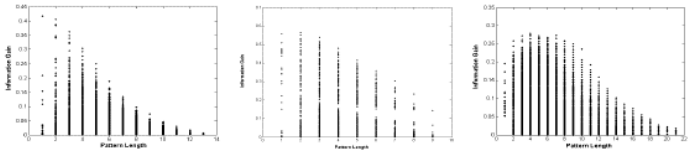
In feature conversion, certain irrelevant features may also be discarded. For example, if in the above case, the colours of the masses $m_1$, & $m_2$ were also provided, it could have been discarded.

## Cheng *et al.*’s Approach




The motivation behind using frequent patterns along with single items is the fact that patterns (up to a certain size) have a much higher “information content”.

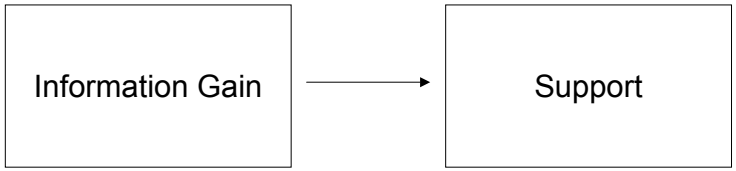
This is demonstrated in the following :



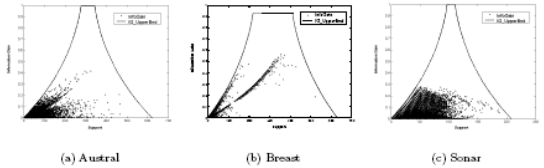
(a) Austral (b) Breast (c) Sonar

The “aim” therefore is to maximize the information content of the feature space.

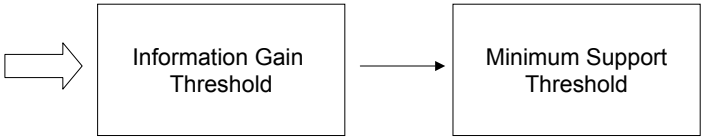
However, the question that arises is : How should the nature of the patterns generated be related to the information content?

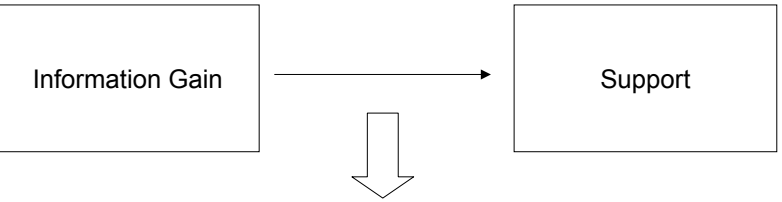



The mapping between the information gain and the support of the data set is clearly visible in the following :




Yang *et al.*[2]




$$IG_{ub}(C \mid X) = H(C) - H_{lb}(C \mid X)$$

$$H(C \mid X) = -\theta q \log q - \theta(1-q)\log(1-q) + (\theta q - p)\log\frac{p-\theta q}{1-\theta}$$
$$+ (\theta(1-q) - (1-p))\log\frac{(1-p)-\theta(1-q)}{1-\theta}$$

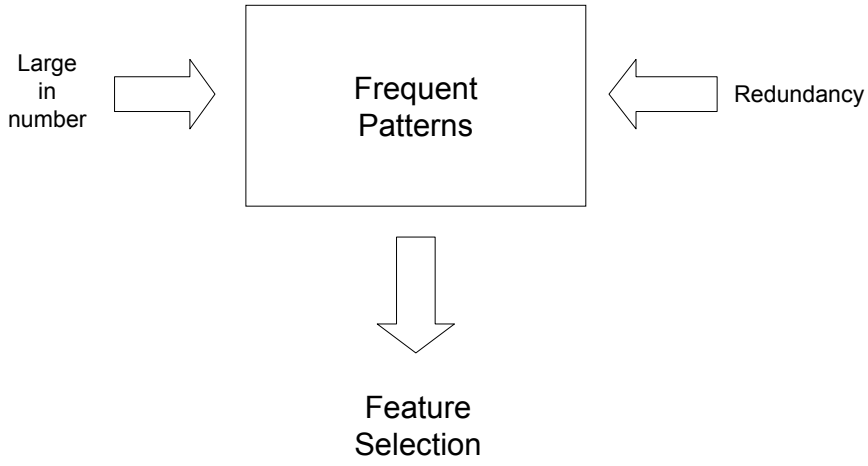
$C = \{0,1\}$

$X \in \{0,1\}$

$P(c=1) = p$

$P(x=1) = \theta$

$P(c=1 \mid x=1) = q$



## Feature Selection

An algorithm (MMRFS) is presented for feature selection.

A few terms that are made use of in the algorithm are defined below.

**Relevance:** ***"A relevance measure S is a function mapping a pattern α to a real value such that S(α) is the relevance w.r.t. the class label."***

**Redundancy :** ***"A redundancy measure R is a function mapping two patterns α and β to a real value such that R(α, β) is the redundancy between them."***

**Gain :** $g(\alpha) = S(\alpha) - \max_{\beta \in F_s} R(\alpha, \beta)$

## Feature Selection Algorithm

**Algorithm 1** Feature Selection Algorithm MMRFS

Input: Frequent patterns $\mathcal{F}$, Coverage threshold $\delta$,
Relevance $S$, Redundancy $R$
Output: A selected pattern set $\mathcal{F}_s$

1: Let $\alpha$ be the most relevant pattern;
2: $\mathcal{F}_s = \{\alpha\}$;
3: while (true)
4: Find a pattern $\beta$ such that the gain $g(\beta)$ is the maximum among the set of patterns in $\mathcal{F} - \mathcal{F}_s$;
5: If $\beta$ can correctly cover at least one instance
6: $\mathcal{F}_s = \mathcal{F}_s \cup \{\beta\}$;
7: $\mathcal{F} = \mathcal{F} - \{\beta\}$;
8: If all instances are covered $\delta$ times or $\mathcal{F} = \phi$
9: break;
10: return $\mathcal{F}_s$

## Experimental Setup

- Datasets from the UCI Machine Learning Repository were made use of.
- Closed frequent patterns were obtained using the FPClose [2] algorithm.
- The MMRFS algorithm was used for feature selection.
- The classifiers used were LIBSVM [3] and C4.5 in Weka [4].
- 10-fold cross-validation was performed to determine the best parameters.

## Experimental Results

| Data | Single Feature | | | Freq. Pattern | |
|---|---|---|---|---|---|
| | *Item_All* | *Item_FS* | *Item_RBF* | *Pat_All* | *Pat_FS* |
| anneal | **99.78** | **99.78** | 99.11 | 99.33 | 99.67 |
| austral | 85.01 | 85.50 | 85.01 | 81.79 | **91.14** |
| auto | 83.25 | 84.21 | 78.80 | 74.97 | **90.79** |
| breast | 97.46 | 97.46 | 96.98 | 96.83 | **97.78** |
| cleve | 84.81 | 84.81 | 85.80 | 78.55 | **95.04** |
| diabetes | 74.41 | 74.41 | 74.55 | 77.73 | **78.31** |
| glass | 75.19 | 75.19 | 74.78 | 79.91 | **81.32** |
| heart | 84.81 | 84.81 | 84.07 | 82.22 | **88.15** |
| hepatic | 84.50 | 89.04 | 85.83 | 81.29 | **96.83** |
| horse | 83.70 | 84.79 | 82.36 | 82.35 | **92.39** |
| iono | 93.15 | 94.30 | 92.61 | 89.17 | **95.44** |
| iris | 94.00 | **96.00** | 94.00 | 95.33 | **96.00** |
| labor | 89.99 | 91.67 | 91.67 | 94.99 | **95.00** |
| lymph | 81.00 | 81.62 | 84.29 | 83.67 | **96.67** |
| pima | 74.56 | 74.56 | 76.15 | 76.43 | **77.16** |
| sonar | 82.71 | 86.55 | 82.71 | 84.60 | **90.86** |
| vehicle | 70.43 | 72.93 | 72.14 | 73.33 | **76.34** |
| wine | 98.33 | 99.44 | 98.33 | 98.30 | **100** |
| zoo | 97.09 | 97.09 | 95.09 | 94.18 | **99.00** |

Accuracy in % obtained using LIBSVM

## Experimental Results

| Dataset | Single Features | | Frequent Patterns | |
|---|---|---|---|---|
| | *Item_All* | *Item_FS* | *Pat_All* | *Pat_FS* |
| anneal | 98.33 | 98.33 | 97.22 | **98.44** |
| austral | 84.53 | 84.53 | 84.21 | **88.24** |
| auto | 71.70 | 77.63 | 71.14 | **78.77** |
| breast | 95.56 | 95.56 | 95.40 | **96.35** |
| cleve | 80.87 | 80.87 | 80.84 | **91.42** |
| diabetes | **77.02** | **77.02** | 76.00 | 76.58 |
| glass | 75.24 | 75.24 | 76.62 | **79.89** |
| heart | 81.85 | 81.85 | 80.00 | **86.30** |
| hepatic | 78.79 | 85.21 | 80.71 | **93.04** |
| horse | 83.71 | 83.71 | 84.50 | **87.77** |
| iono | 92.30 | 92.30 | 92.89 | **94.87** |
| iris | **94.00** | **94.00** | 93.33 | 93.33 |
| labor | 86.67 | 86.67 | **95.00** | 91.67 |
| lymph | 76.95 | 77.62 | 74.90 | **83.67** |
| pima | 75.86 | 75.86 | 76.28 | **76.72** |
| sonar | 80.83 | 81.19 | **83.67** | **83.67** |
| vehicle | 70.70 | 71.49 | **74.24** | 73.06 |
| wine | 95.52 | 93.82 | 96.63 | **99.44** |
| zoo | 91.18 | 91.18 | 95.09 | **97.09** |

Accuracy in % obtained using C4.5

## Experimental Results (scalability)

| *min_sup* | #Patterns | Time (s) | SVM (%) | C4.5 (%) |
|---|---|---|---|---|
| 1 | N/A | N/A | N/A | N/A |
| 2000 | 68,967 | 44.703 | 92.52 | 97.59 |
| 2200 | 28,358 | 19.938 | 91.68 | 97.84 |
| 2500 | 6,837 | 2.906 | 91.68 | 97.62 |
| 2800 | 1,031 | 0.469 | 91.84 | 97.37 |
| 3000 | 136 | 0.063 | 91.90 | 97.06 |

Chess Data

-

| *min_sup* | #Patterns | Time (s) | SVM (%) | C4.5 (%) |
|---|---|---|---|---|
| 1 | 9,468,109 | N/A | N/A | N/A |
| 80 | 26,576 | 176.485 | 92.40 | 88.35 |
| 100 | 15,316 | 90.406 | 92.19 | 87.29 |
| 150 | 5,408 | 23.610 | 91.53 | 88.80 |
| 200 | 2,481 | 8.234 | 91.22 | 87.32 |

Waveform Data

## Experimental Results (scalability)

| *min_sup* | #Patterns | Time (s) | SVM (%) | C4.5 (%) |
|---|---|---|---|---|
| 1 | 5,147,030 | N/A | N/A | N/A |
| 3000 | 3,246 | 200.406 | 79.86 | 77.08 |
| 3500 | 2,078 | 103.797 | 80.21 | 77.28 |
| 4000 | 1,429 | 61.047 | 79.57 | 77.32 |
| 4500 | 962 | 35.235 | 79.51 | 77.42 |

Letter Recognition Data

## Critique

- The justification provided for each step in the approach adopted is commendable.
- The experiments have been conducted comprehensively.
- The method introduced is more of a formalization of methods already in use, rather than an entirely new approach.
- Certain parameters and terms used have not been explained nor provided with adequate citations.

## References


1. H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative feature pattern analysis for effective classification. (submitted)

2. Y.Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of ICML,* pages 412-420, 1997.

3. C.-C. Chang and C.-J. Lin. *LIBSVM: a library of support vector machines,* 2001. Software available at http://www.csei.nyu.edu.tw/~cjlin/libsvm.

4. I. H. Witten and E. Frank. *Data Mining; Practical machine learning tools and techniques.* Morgan Kaufmann, 2nd edition, 2005.


Thank you!