Feature Selection with Labelled and Unlabelled Data

Shaomin Wu and Peter A. Flach

Department of Computer Science, University of Bristol,
Woodland Road, Bristol BS8 1UB, U.K
{shaomin,flach}@cs.bris.ac.uk

Abstract. Most feature selection approaches perform either exhaustive or heuristic search for an optimal set of features. They typically only consider the labelled training set to obtain the most suitable features. When the distribution of instances in the labelled training set is different from the unlabelled test set, this may result in large generalization error. In this paper, a combination of heuristic measures and exhaustive search based on both the labelled dataset and the unlabelled dataset is proposed. The heuristic measures concerned are two contingency table measures — Goodman-Kruskal measure and Fisher's exact test — which are used to rank the feature according to how well a feature predicts the class. Secondly, an exhaustive search is employed: by using test for goodness-of-fit, information on both the labelled dataset and the unlabelled dataset is applied to choose a better combination of features. We evaluate the approaches on the KDD Cup 2001 dataset.

1. Introduction

Feature selection aims at finding a feature subset that can describe the data for a learning task as good as or better than the original dataset. It is of importance for both data mining and machine learning, in particular for high-dimensional data. Most algorithms for feature selection perform either heuristic or exhaustive search [1]. Heuristic feature selection algorithms estimate the feature's quality with a heuristic measure, for instance, information gain [2], Gini index [3], discrepancies measure [4] and chi-square test [5]. Other examples of heuristic algorithms include the Relief algorithm [6] and its extension, the PRESET algorithm [7]. Exhaustive feature selection algorithms search all possible combinations of features and aim at finding a minimal combination of features that is sufficient to construct a model consistent with a given set of instances, for example, the FOCUS algorithm [8].

In supervised learning we use a labelled training set to obtain a model, which is then executed on an unlabelled test set to obtain predictions. However, the model developed from the labelled dataset may not perform well on prediction for the unlabelled dataset because of differences in class distribution and cost distribution between the labelled dataset and the unlabelled data. For classification, ROC analysis [9] can be used to choose the best model from a model set if distribution of positives and negatives and distribution of misclassification costs for the unlabelled dataset are given. Whereas misclassification costs may be given, it is often impossible to obtain the class distribution of positives and negatives for the unlabelled data. What can be obtained from the unlabelled data is information about the distribution of instances. For instance, the transduction technique [10] aims at maximizing the classification margin on both the labelled and the unlabelled data.

Algorithms for both heuristic and exhaustive feature selection in the literature, however, only focus on the labelled dataset. This may lead to large generalization error when the instance distribution in the labelled dataset is different from that of the unlabelled data. This paper introduces two feature selection approaches: feature selection based on the Goodman-Kruskal measure and feature selection based on both labelled and unlabelled datasets. The Goodman-Kruskal measure is used to select a subset of features, which is then exhaustively searched for a sub-subset with similar distributions in both the labelled and the unlabelled datasets. Experimental evaluation shows that the proposed approach performs well compared with other feature selection approaches.

The paper is organized as follows. Section 2 introduces two contingency table measures — the Goodman-Kruskal measure and Fisher's exact test — to rank the importance of features. Section 3 proposes a new feature selection approach based on the unlabelled dataset and the Chi-squared test for goodness-of-fit. Section 4 evaluates the approach on the KDD Cup 2001 dataset [11]. Section 5 concludes with a discussion and the main conclusions.

2. Heuristic measures for feature selection

Heuristic feature selection algorithms search the feature set with a heuristic measure such as information gain. Assume that the input features are independent of each other, we can compare associations between each input feature and the class to select important features. Let the value of the class be P (positive) or N (negative), and the value of an input feature be $C_1\ldots, C_r,$ then a contingency table can be built up as follows.

	Class = P	Class = N	
Input Feature = C_1	n_{1P} (μ_{1P})	n_{1N} (μ_{1N})	n_{1*}
…	…	…	…
Input Feature = C_r	n_{rP} (μ_{rP})	n_{rN} (μ_{rN})	n_{r*}
	n_{*P}	n_{*N}	n

Table 1. An $r \times 2$ contingency table

In table 1, n_{ij} is the number of instances for which the value of a feature is C_i and the value of the class is j. $n_{*j} = \sum_{i=1}^{r} n_{ij}$, $n_{i*} = n_{iP} + n_{iN}$, $n = n_{*P} + n_{*N}$, and

$\mu_{ij} = \frac{n_{*j} n_{i*}}{n}$, where n is the number of instances in the labelled data, $i = 1, \cdots, r$ and j=P, N. The table has r–1 degrees of freedom.

2.1 Chi-squared measure

Most methods to measure the association between two features in a contingency table are based on the Chi-squared test [5]. The Chi-squared measure can be used to measure the association between class and input feature. In the 2-by-r case in Table 1 it is defined as

$$c^2 = \sum_{i=1}^{r} \left(\frac{(n_{iP} - m_{iP})^2}{m_{iP}} + \frac{(n_{iN} - m_{iN})^2}{m_{iN}} \right) \tag{1}$$

This value is compared with a threshold value corresponding to a confidence level. For instance, if r=2 (1 degree of freedom) the χ^2 value at the 5% level is 3.84 — if our χ^2 value is larger than that, the probability is less than 5% that discrepancies this large are attributable to chance, and we are led to reject the null hypothesis of independence.

2.2 Fisher's exact measure

If one uses the Chi-squared measure to test whether an association exists between two random variables, m_{ij}>5 should be satisfied for each i and j. When $m_{ij} \leq 5$ and r=2, Fisher's exact test can be applied to test the association. Assume that $m_{1P} \leq 5$, $n_{1*} \leq n_{*P}$ and $n_{1*} \leq n_{*N}$. Below is Fisher's exact measure [13]

$$P_F = \sum_{k=n_{1P}}^{n_{1*}} \frac{n_{1*}!n_{2*}!n_{*P}!n_{*N}!}{k!(n_{1*} - k)!(n_{*P} - k)!(n_{*N} - n_{1*} + k)!(n_{1*} + n_{2*})!} \tag{2}$$

This measure is normalised between 0 and 1. When P_F is less than 0.05, we are led to reject the null hypothesis of independence (at the 5% level). It should be noted that Fisher's exact test can become computationally expensive for large n and r.

2.3 Goodman-Kruskal measure

The Chi-squared measure and Fisher's exact test can only measure the association between two features, as they are symmetric in the two features. Goodman and Kruskal [14] introduced an asymmetric measure λ that measures the predictivity of one feature with respect to another, say, predicting class with an input feature. The measure is

$$l = \frac{\sum_{i=1}^{r} \max(n_{iP}, n_{iN}) - \max(n_{*P}, n_{*N})}{n - \max(n_{*P}, n_{*N})} \quad (3)$$

where $0 \le \lambda \le 1$. λ=0 means no predictive gain when using an input feature to predict the class, and λ=1 means perfect predictivity. If we want to select features that have strong association with the class in a dataset, both n_{*P} and n_{*N} which are the number of instances in which the class equals to P and N, respectively, will be constant. In this case, a simplified version of the Goodman-Kruskal measure is

$$\lambda_0 = \sum_{i=1}^{r} \max(n_{iP}, n_{iN}) \quad (4)$$

Section 4 in this paper will give examples that performance of models based on features selected with Goodman-Kruskal measure is sometimes better than those based on Chi-squared measure and information gain measure.

2.4 Information gain

For the sake of comparison, we use a feature selection approach based on information gain. Information gain is commonly used as a surrogate for approximating a conditional distribution in the classification setting [15]. Below is a simplified version of information gain for our problem (the remaining part only depends on the class).

$$I_{gain} = -\sum_{i=1}^{r} \left(\frac{n_{iP}}{n_{1*}} \log \frac{n_{iP}}{n_{1*}} + \frac{n_{iN}}{n_{2*}} \log \frac{n_{iN}}{n_{2*}} \right) \quad (5)$$

3. Feature selection based on labelled and unlabelled data

Heuristic measures like the above can be used to rank features. However, such a ranking does not consider that the probability distribution of the features in the labelled dataset may be different from those in the unlabelled dataset. In general, a model developed from the labelled dataset may have large generalization error if the probability distribution of the model's features in the labelled dataset is considerably different from their distribution in the unlabelled dataset. Assume M features, say, $x_1, x_2, \cdots, x_M$, are selected with a certain criterion from N features, where $M \le N$, and a model below is built up based on the M features.

$$y = f(x_1, x_2, \cdots, x_M) \quad (6)$$

where y represents the class. The probability distribution of $x_1, x_2, \cdots, x_M$ in the labelled dataset is expected to be close to the one in the unlabelled data. In other words, the closer the probability distributions of $x_1, x_2, \cdots, x_M$ between the labelled dataset and the unlabelled data, the lower generalization error the model (6) has.

Let the probability distribution of $x_1, x_2, \cdots, x_M$ in the labelled dataset be $F(x_1, x_2, \cdots, x_M)$. According to the assumption of the heuristic search, $x_1, x_2, \cdots, x_M$ are independent of each other. Therefore, $F(x_1, x_2, \cdots, x_M)$ can be simplified as

$$F(x_1, x_2, \cdots, x_M) = F_1(x_1) F_2(x_2) \cdots F_M(x_M) \tag{7}$$

where $F_i(x_i)$ $(i = 1, 2, \cdots, M)$ is the probability distribution of x_i in the labelled data. Similarly, let the probability distribution of $x_1, x_2, \cdots, x_M$ in the unlabelled dataset be $G(x_1, x_2, \cdots, x_M)$, we have

$$G(x_1, x_2, \cdots, x_M) = G_1(x_1) G_2(x_2) \cdots G_M(x_M) \tag{8}$$

where $G_i(x_i)$ $(i = 1, 2, \cdots, M)$ is the probability distribution of x_i in the unlabelled data.

If the distribution function $F(x_1, x_2, \cdots, x_M)$ and $G(x_1, x_2, \cdots, x_M)$ come from the same distribution, the performance of the model on the labelled dataset and on the unlabelled dataset will be similar. If the probability distributions $F_i(x_i)$ and $G_i(x_i)$ are similar, the distribution functions $F(x_1, x_2, \cdots, x_M)$ and $G(x_1, x_2, \cdots, x_M)$ will be similar.

Assuming that x_i is a categorical feature, the Chi-squared test for goodness-of-fit can be used to estimate whether two random variables come from the same distribution. For the labelled data, let π_{ij} be the probability that the value of feature i falls in category C_{ij}, $j = 1, 2, \ldots, C$, which can be estimated as

$$\pi_{ij} = \frac{\text{the number of instances with feature } i \text{ falling in category } C_{ij} \text{ in labelled data}}{\text{the total number of instances in training dataset}} \tag{9}$$

Similarly for the unlabelled data, let θ_{ij} be the probability that the value of feature i falls in category C_{ij}, estimated as

$$\theta_{ij} = \frac{\text{the number of instances with feature } i \text{ falling in category } C_{ij} \text{ in unlabelled data}}{\text{the total number of instances in working dataset}} \tag{10}$$

The Chi-squared statistic for goodness-of-fit can be employed to measure the closeness between the distribution π_{ij} and θ_{ij} for $j = 1, \ldots, C$:

$$\chi_i^2 = n \sum_{j=1}^{C} \frac{(\theta_{ij} - \pi_{ij})^2}{\pi_{ij}} \tag{11}$$

The smaller the value of χ_i^2 is, the more similar the distributions π_{ij} and θ_{ij} are. We average this over all features as follows:

$$c_{new} = \frac{1}{M}\sum_{i=1}^{M} c_i^2 \tag{12}$$

which measures the similarity between $F(x_1, x_2, \cdots, x_M)$ and $G(x_1, x_2, \cdots, x_M)$.

We can now formulate our proposed feature selection approach. We first use a heuristic measure to select the N^* best features, then apply a exhaustive search based on measure χ_{new} to select the combination of M features which minimizes the value of χ_{new}, where $M < N^* < N$.

4. Experimental Evaluation

The thrombin dataset from KDD Cup 2001 consists of 139351 features and 1909 instances and one class in the labelled data. All features and the class are binary. There are 42 instances labelled ‘A’ (standing for ‘active’, the positive class) and 1867 instances labelled ‘I’ (standing for ‘inactive’, the negative class). Below is the contingency table.

	Class Activity =A	**Class Activity =I**	
Input Feature = ’1’	$n_{1A}(\mu_{1A})$	$n_{1I}(\mu_{1I})$	n_{1*}
Input Feature = ‘0’	$n_{0A}(\mu_{0A})$	$n_{0I}(\mu_{0I})$	n_{0*}
	42	1867	1909

Table 2. Contingency table.

A test dataset (below we call it the unlabelled data) with 634 unlabelled instances is given. We will use this dataset to test the performance of models. ROC analysis is used compare the performances of several classifiers within a ROC space. It allows, through the construction of the convex hull of a set of points, identification of classifiers that are optimal under certain parameter settings. Once the application context is known, say, distribution of positives and negatives and misclassification costs, the optimal classifiers can be determined from the convex hull. Only the classifiers on the convex hull are optimal under some circumstances.

4.1 Heuristic feature selection

Chi-squared measure, Fisher’s exact measure, Goodman-Kruskal measure and information gain measure discussed above are used to select the features.

If the measure χ^2 in equation (1) is used to select features, 120941 features can be selected from the labelled dataset when a criterion $\chi^2 > 3.84$ is applied.

When $n_{1*} < 228$, μ_{1A} will be less than 5. Chi-squared test will not be suitable for the case and Fisher's exact measure P_F in section 2.2 can be used. In order to simplify the calculation, Fisher's exact test P_F in equation (2) is applied no matter whether n_{1*} is greater than or less than 228. 102326 features whose P_F values are all less than 0.05 can be selected.

By using Goodman-Kruskal measure in equation (4), we can select 51540 features whose λ are greater than zero.

If information gain measure in equation (5) is used to select the features, a set of features with ascending order of measure can be obtained. The set only shows the importance of each feature.

At the same time, the ID3 algorithm is used to build a decision based on the whole labelled dataset, and eight features occur in the tree.

Because we only focus on the comparison of different measures instead of the number of features to be selected, in order to compare and simplify our calculation, analogous to the number of features selected by the ID3 decision tree, eight features selected with other measures are chosen to develop models.

As it turns out, the first eight features selected with Fisher's exact measure are the same as those selected with the information gain measure. The order of features selected with Fisher's exact measure and information gain measure is very similar. Unfortunately, it is hard to prove that measure of P_F in equation (2) and the measure of I_{gain} in equation (5) can lead a similar result. Because the features selected by information gain and those by Fisher's exact measure are the same, we only compare the information gain measure with other measures below.

Based on the labelled data, logistic regression model, Naive Bayes model, IB1 model, support vector machine---sequential minimal optimisation (SMO) algorithm, Kstar model and ID3 decision tree are built. Descriptions of those approaches can be found in [12]. The Weka toolkit is used to build the models. We use ten-fold cross validation to estimate the error of a model on the labelled data. In order to explain the evaluation metrics used, let the confusion matrix be

	Positive examples	Negative examples
Instances predicted positive	a	b
Instances predicted negative	c	d

Table 3. A confusion matrix

Then the metrics are as follows:

Accuracy=$\dfrac{a+d}{a+b+c+d}$,

Recall Average=$\dfrac{1}{2}(\dfrac{a}{a+c}+\dfrac{d}{b+d})$, True Positive Rate= $\frac{a}{a+c}$, and False Positive Rate= $\frac{c}{b+d}$.

A model is expected to possess high accuracy, high recall average, high TPrate and low FPrate.

In Table 4, for instance, 0.833(0.991) in the second row and the second column means, recall average=0.991 and accuracy=0.833 for the logistic model on the labelled dataset (lbl). 0.484(0.416) in the third row and the second column means, recall average=0.484 and accuracy=0.416 for the logistic model on the unlabelled dataset (ulbl), and so on. ID3 in the first row means the feature set selected by the ID3 decision tree, Info Gain means information gain measure, Chi-squared means Chi-squared measure and Goodman means Goodman-Kruskal measure.

The underlined numbers indicate the maximum value in the same row in table 4. From the table, both accuracy and recall average are maximum for Naive Bayes model and support vector machine and accuracy for Prism model and ID3 decision tree are maximum for the unlabelled dataset when Goodman-Kruskal measure is used, and no performance for other measures is better than Goodman-Kruskal measure.

Model	ID3	Info Gain	Chi Squared	Goodman
Logistic (lbl)	0.833(0.991)	0.736(0.984)	0.749(0.986)	0.772(0.985)
Logistic (ulbl)	0.484(0.416)	0.509(0.626)	0.564(0.598)	0.546(0.597)
NaiveBayes (lbl)	0.838(0.979)	0.839(0.982)	0.820(0.988)	0.900(0.988)
NaiveBayes (ulbl)	0.480(0.402)	0.552(0.379)	0.538(0.435)	0.543(0.544)
SMO (lbl)	0.806(0.984)	0.806(0.984)	0.808(0.988)	0.819(0.987)
SMO (ulbl)	0.485(0.319)	0.500(0.457)	0.525(0.509)	0.564(0.615)
Prism (lbl)	0.812(0.989)	0.682(0.983)	0.756(0.987)	0.707(0.985)
Prism (ulbl)	0.501(0.497)	0.492(0.655)	0.577(0.587)	0.607(0.576)
ID3 (lbl)	0.867(0.990)	0.748(0.985)	0.808(0.988)	0.748(0.985)
ID3 (ulbl)	0.495(0.475)	0.542(0.642)	0.609(0.697)	0.618(0.651)
IB1 (lbl)	0.808(0.988)	0.795(0.986)	0.761(0.987)	0.748(0.985)
IB1 (ulbl)	0.548(0.555)	0.551(0.662)	0.541(0.584)	0.533(0.596)
Kstar (lbl)	0.702(0.986)	0.724(0.985)	0.774(0.990)	0.737(0.988)
Kstar (ulbl)	0.449(0.569)	0.534(0.632)	0.560(0.623)	0.544(0.645)

Table 4. Results based on different measures

Figure 1 is a ROC curve based on models for the unlabelled dataset. The X-axis and the Y-axis in the ROC curve represent FPrate and TPrate, respectively. The ROC curve shows that Naïve Bayes model based on features selected with information gain measure, both Prism model and ID3 decision tree based on features selected with Goodman-Kruskal measure and ID3 decision tree based on features selected with Chi-squared measure are on the convex hull. In other words, those models are the optimal models in certain circumstances.

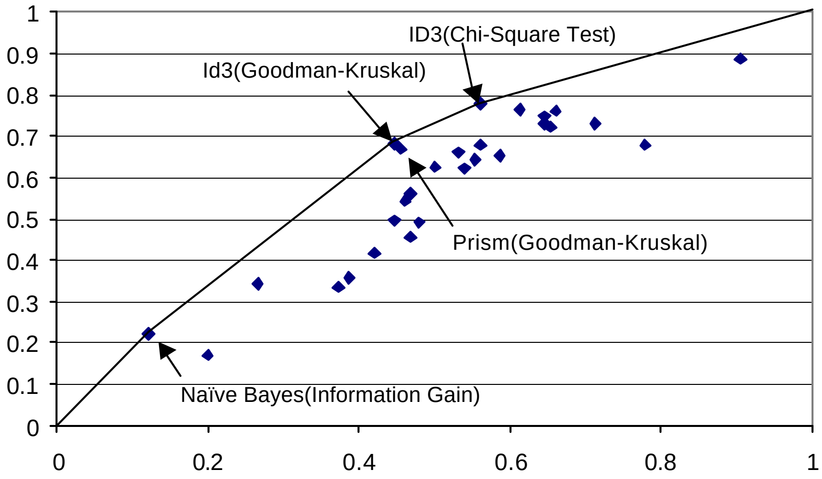

Fig. 1. ROC curve---comparison of different measures

4.2 Feature selection based on labelled and unlabelled data

By using Goodman-Kruskal measure, all features in the labelled dataset can be ranked in a descending order. The first twenty and thirty features from the ranked feature set are selected to form two feature sets, respectively. Then an exhaustive search for this two sets is performed to find the best combination of eight features to minimize χ_{new} in equation (12). Let F20 and F30 be the exhaustive search on the twenty-feature set and on the thirty-feature set, respectively. Table 5 shows the performances of models based on features selected with different measures.

The underlined numbers indicate the maximum values in the same row for the unlabelled dataset in table 5. From the table, only Naïve Bayes model based on features selected with information gain measure and ID3 decision tree based on features with Goodman-Kruskal measure have higher accuracy than F20 search or F30 search. Logistic model based on features selected with information gain measure, support vector machine model based on features selected with Goodman-Kruskal measure and ID3 decision tree model based on features selected with Chi-squared measure have higher recall average than F20 search or F30 search.

Figure 2 is a ROC curve based on prediction of models for the unlabelled dataset. The ROC curve shows that Naïve Bayes model based on features selected with information gain measure, Prism model on features selected with F20 are on the convex hull, IB1 on features selected with F20, Prism model on features selected with F30 and Kstar-kstar model on features selected with F30 are close to convex hull, which mean that F20 and F30 are better than other feature selection approaches.

Model	ID3	Info Gain	Chi Squared	Goodman	F20	F30
Logistic (lbl)	0.833(0.991)	0.736(0.984)	0.749(0.986)	0.772(0.985)	0.772(0.985)	0.784(0.987)
Logistic (ulbl)	0.484(0.416)	0.509(0.626)	0.564(0.598)	0.546(0.597)	0.559(0.623)	0.588(0.691)
NaiveBayes (lbl)	0.838(0.979)	0.839(0.982)	0.820(0.988)	0.900(0.988)	0.785(0.990)	0.820(0.990)
NaiveBayes (ulbl)	0.480(0.402)	0.552(0.379)	0.538(0.435)	0.543(0.544)	0.514(0.569)	0.579(0.691)
SMO (lbl)	0.806(0.984)	0.806(0.984)	0.808(0.988)	0.819(0.987)	0.761(0.986)	0.761(0.987)
SMO (ulbl)	0.485(0.319)	0.500(0.457)	0.525(0.509)	0.564(0.615)	0.604(0.577)	0.576(0.696)
Prism (lbl)	0.812(0.989)	0.682(0.983)	0.756(0.987)	0.707(0.985)	0.723(0.987)	0.774(0.989)
Prism (ulbl)	0.501(0.497)	0.492(0.655)	0.577(0.587)	0.607(0.576)	0.688(0.735)	0.614(0.722)
ID3 (lbl)	0.867(0.990)	0.748(0.985)	0.808(0.988)	0.748(0.985)	0.738(0.988)	0.773(0.988)
ID3 (ulbl)	0.495(0.475)	0.542(0.642)	0.609(0.697)	0.618(0.651)	0.644(0.667)	0.610(0.721)
IB1 (lbl)	0.808(0.988)	0.795(0.986)	0.761(0.987)	0.748(0.985)	0.785(0.988)	0.773(0.988)
IB1 (ulbl)	0.548(0.555)	0.551(0.662)	0.541(0.584)	0.533(0.596)	0.677(0.744)	0.569(0.689)
Kstar (lbl)	0.702(0.986)	0.724(0.985)	0.774(0.990)	0.737(0.988)	0.726(0.988)	0.691(0.986)
Kstar (ulbl)	0.449(0.569)	0.534(0.632)	0.560(0.623)	0.544(0.645)	0.571(0.716)	0.561(0.751)

Table 5. Comparisons of different approaches.

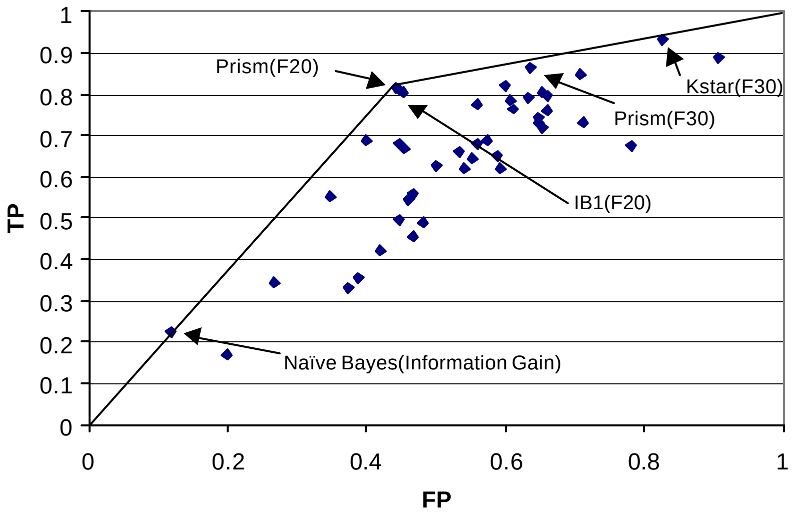

Fig. 2. ROC curve comparing different measures based on labelled and unlabelled data.

5. Concluding remarks

In the above experiments in section 4, we first selected candidate feature set (say, twenty and thirty features in this paper) from features selected with Goodman-Kruskal measure, then use χ_{new} in equation (13) to search a best combination of eight features among the candidate feature set. However, if the candidate feature set is too large, some features with small Goodman-Kruskal measure may be included. Therefore, if more features are searched with χ_{new}, a smaller χ_{new} may probably be achieved, but prediction association between features and the class will degrade which means the performance of models based on features selected with χ_{new} will become poor. If few features are searched with χ_{new}, a larger χ_{new} is probably obtained. That means that the distribution of the labelled dataset and that of the unlabelled dataset may have a big difference, which will lead to larger generalization error. In other words, there is a trade-off between the size of search space and the value of χ_{new}.

Contingency table measures have been discussed by statisticians for a long time. The most well-known technique for analyzing the contingency table is the Chi-squared test. Furthermore, Fisher's exact test is used to test on contingency table with small expectations and Goodman-Kruskal measure is used to measure the prediction association. This paper firstly borrowed Fisher's exact test, Goodman-Kruskal measure to select feature. Below we summarize the results given in this paper

A. The rank order with Fisher's exact measure is similar to the one with information gain measure.
B. The performance of feature selection based on the Goodman-Kruskal measure is better than those based on other measures.
C. Feature selection with a measure based on the features from the labelled dataset and the unlabelled dataset has a lower generalization error than those based only on the labelled dataset.

Acknowledgements

This work is supported by the Esprit V project (IST-1999-11495) *Data Mining and Decision Support for Business Competitiveness: Solomon Virtual Enterprise*. We thank DuPont Pharmaceuticals Research Laboratories and KDD Cup 2001 for graciously providing this data set. Thanks are also due to the anonymous reviewers for their insightful comments and suggestions.

Reference:

[1] Blum, A., Langley, P., *Selection of relevant features and examples in machine learning*. Artificial intelligence, 97, 1997, pp.245-271

[2] Hunt, E., martin, J., and Stone, P. *Experiments in Induction*. Academic Press, New York, 1966

[3]Breiman, L., et. al., *Classification and regression trees*. Wadsworth Inc., Belmont, California, 1984

[4] Mantara, R., *ID3 revisited: A discrepancies based criterion for attribute selection*. In Proceedings of International Symposium Methodologies for Intelligent Systems, Charlotte, North Carolina, USA, 1989

[5] Lehmann, E. L., *Testing statistical hypothesis*. Springer, 199

[6] Kira, K., and Rendell, L., *The feature selection problem: Traditional methods and a new algorithm*. In Proceedings of the tenth National Conference on Artificial intelligence. Menlo Park: AAAI Press/The MIT Press, 1992, pp. 129-134

[7] Modrsejewski, M., *Feature selection using rough sets theory*, in P.B.Brazdil, ed., Proceedings of the European Conference on Machine Learning, Springer, 1993, pp. 213-226

[8]Almuallim, H., and Dietterich, T., *Learning boolean concepts in the presence of many irrelevant feature*. Artificial Intelligence, 69(1-2), 1994, pp. 279-305

[9] Provost, F., and Fawcett, T., *Robust Classification for Imprecise Environments, Machine Learning*, 42, 2001, pp. 203–231.

[10] Vapnik, V. N., *Statistical Learning Theory*. Wiley, 1998.

[11] Page, D, www.cs.wisc.edu/~dpage/kddcup2001/

[12] Witten, I. H., and Frank E., *Data Mining -Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000

[13] Goodman, L. A. and Kruskal, W. H., *Measures of association for classifications*. J. Amer. Statist. Ass. 49, 1954, pp.732-764

[14] Agresti, A., *A Survey of Exact Inference for Contegency Tables.* Statitical Science, **7**, 1992, pp.131-153

[15] Cover, T. and Thomas, J., *Elements of Information Theory*, Wiley, 1991