T.C.

BAHÇEŞEHİR ÜNİVERSİTESİ

A NOVEL GENERALIZED MUTUAL INFORMATION APPROACH AND ITS USE IN FEATURE SELECTION

Master Thesis

CEMAL OKAN ŞAKAR

İSTANBUL, 2008

T.C.

BAHÇEŞEHİR ÜNİVERSİTESİ

INSTITUTE OF SCIENCE

COMPUTER ENGINEERING

A NOVEL GENERALIZED MUTUAL INFORMATION APPROACH AND ITS USE IN FEATURE SELECTION

Master Thesis

CEMAL OKAN ŞAKAR

Supervisor: ASST. PROF. DR. OLCAY KURŞUN

İSTANBUL, 2008

T.C.

BAHÇEŞEHİR ÜNİVERSİTESİ

INSTITUTE OF SCIENCE

COMPUTER ENGINEERING

Name of the thesis: A Novel Generalized Mutual Information Approach and Its Use In Feature Selection
Name/Last Name of the Student: Cemal Okan ŞAKAR
Date of Thesis Defense: 03 June 2008

The thesis has been approved by the Institute of Science.

Prof. Dr. Erol SEZER
Director

____

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Adem KARAHOCA
Program Coordinator

____

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members	Signature
Asst. Prof. Dr. Olcay KURŞUN	____________________
Prof. Dr. Emin ANARIM	____________________
Assoc. Prof. Dr. Adem KARAHOCA	____________________

ACKNOWLEDGMENTS

This thesis is dedicated to **my respectable grandfather, Bekir KERTMEN**;
and to **my family**.

I would like to express my gratitude to my supervisor **Asst. Prof. Dr. Olcay KURŞUN** for encouraging and challenging me throughout my thesis studies.

I also thank **Prof. Dr. Şenay YALÇIN** for his helps on various topics throughout my academic program and also my life.

ABSTRACT

A NOVEL GENERALIZED MUTUAL INFORMATION APPROACH
AND
ITS USE IN FEATURE SELECTION

ŞAKAR, Cemal Okan

Computer Engineering

Supervisor: Asst. Prof. Dr. Olcay KURŞUN

June 2008, 50 Pages

Feature selection is a critical step in many artificial intelligence and pattern recognition problems. Shannon's Mutual Information (*MI*) is a classical and widely used measure of dependence measure that serves as a good feature selection algorithm. However, as it is a measure of mutual information in average, under-sampled classes (rare events) can be overlooked by this measure, which can cause critical false negatives (missing a relevant feature very predictive of some rare but important classes). Shannon's mutual information requires a well sampled database, which is not typical of many fields of modern science (such as biomedical), in which there are only a limited number of samples to learn from, or at least, not all the classes of the target function (such as certain phenotypes in biomedical) are well-sampled. Moreover in such settings, each feature, among many, contributes in small amounts to the target function to be predicted, analyzed, or modeled. A new measure of relevance, Predictive Mutual Information (*PMI*), is proposed in this thesis which also accounts for predictability of signals from each other in its calculation. *PMI* has more improved feature detection capability than *MI*, especially in catching suspicious coincidences that are rare but potentially important not only for experimental studies but also for building computational models. This measure, in its formulation, turns out to be a generalization of Shannon's mutual information. Moreover, *PMI* is further developed with the aim of selecting the most compact set of most relevant variables (with minimal redundancies among them). The usefulness of *PMI* and superiority over *MI* is demonstrated on both toy and real datasets.

Keywords: Suspicious Coincidences; Statistical Dependence; Under Sampling; Classification and Inferential Models; Data Mining and Visualization.

ÖZET

YENİ BİR GENELLEŞTİRİLMİŞ KARŞILIKLI BİLGİ YAKLAŞIMI
VE
DEĞİŞKEN SEÇİMİNDE KULLANIMI

ŞAKAR, Cemal Okan

Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç. Dr. Olcay KURŞUN

Haziran 2008, 50 Sayfa

Değişken seçimi birçok yapay zeka ve örnek tanıma problemlerinin kritik adımlarından biridir. Shannon'ın karşılıklı bilgi (*KB*) ölçümü iyi bir değişken seçim algoritması olarak yaygın şekilde kullanılmaktadır. Ancak *KB* ortalama karşılıklı bilgiyi iyi ölçmesine rağmen, örnek sayısı az olan sınıfları (ender olayları) gözden kaçırarak yanlış sınıflandırmalara neden olabilmektedir (önemli ama ender rastlanan bu sınıflar hakkında bilgi içeren alakalı değişkenlerin kaçırılması sonucunda). *KB* iyi örneklenmiş veri kümelerine ihtiyaç duyar; bu da özellikle biomedikal alanındaki gibi sınırlı sayıda örneği olan veya en azından, bazı sınıfları iyi örneklenmemiş (biomedikal alanında ender rastlanan hastalık, kanser örnekleri gibi) veri kümelerine sahip modern bilim dallarında kullanımını verimsizleştirir. Ayrıca bu tip veri kümelerinde değişkenler, tahmin, analiz ve modelleme yapılacak hedef değişkene ancak küçük katkılar yapar. Bu tez çalışmasında, değişkenlerin kendi aralarındaki koşullu olasılıklarını da dikkate alan yeni bir istatistiksel ilişki metriği, Koşullu Karşılıklı Bilgi (*KKB*), önerilmiştir. *KKB*, *KB*’ye kıyasla, sadece deneysel çalışmalarda değil, bilgisayar ile işaret tanıma modellerinin oluşturulmasında da önemli olan şüpheli derecede ilginç durumları yakalamada daha başarılı değişken seçebilmektedir. Bu metrik, formülasyonu itibariyle *KB*’nin bir genel halidir. Buna ek olarak, *KKB*’yi, aralarında ortak bilgi taşıyan değişkenleri mümkün olduğu kadar az seçecek şekilde daha da geliştirerek, mümkün olan en az sayıda ama hedef değişken ile azami karşılıklı bilgi içereek bir değişken seçimi metodu önerilmiştir. *KKB*’nin kullanışlılığı ve *KB*’ye olan üstünlükleri yapay ve gerçek veri kümeleri üzerinde gösterilmiştir.

Keywords: Şüpheli Derecede İlginç Durumlar; İstatistiksel İlişki Metrikleri; Sınırlı Örnekleme; Sınıflandırma ve Tahmin Yürütme Modelleri; Veri Madenciliği and Veri Görüntüleme.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Artificial Neural Networks	:	ANN
Kullback – Leibler	:	KL
Maximum Relevance Minimum Redundancy	:	mRMR
Multilayer Perceptron	:	MLP
Mutual Information	:	*MI*
Predictive Mutual Information	:	*PMI*
Probability Density Function	:	PDF
Radial Basis Function	:	RBF
Shannon’s Mutual Information	:	*MI*
Sulfur Dioxide	:	SO_2
Support Vector Machines	:	SVM

LIST OF SYMBOLS

Activity of a hidden unit in an MLP	: H_h
Activity of the output unit in an MLP	: y
Correlation Coefficient	: ρ
Joint entropy of *X* and *Y*	: $H(X,Y)$
Modified joint probability distribution	: Q'
Mutual information between *X* and *Y*	: $I(X; Y)$
Normalized mutual information	: $N(X;Y)$
Predictive mutual information between *X* and *Y*	: $PMI(X;Y)$
Probability distribution function of *X*	: $P(X)$
Shannon's Entropy	: $H(X)$
The error signal in an MLP	: δ
Weight decay constant	: γ

1. INTRODUCTION

Modern science turns to progressively more complex and challenging subjects across many fields – medicine, neuroscience, genomics and related fields, ecology, economics, climatology, cosmology, etc. This expansion of scientific inquiry into until recently inaccessible territories is brought about by ever growing advances in computer and sensor technologies, which enable the collection of large amounts of groundbreaking novel experimental and observational data. On the other hand, the new subjects also address more complex phenomena, in which many factors contribute to the target function but in small amounts (Ioannidis 2005). Therefore, feature selection stands as one of the major problems in machine learning and pattern recognition applications. The goal of feature selection is to choose the most relevant subset of variables among many, thus reducing the dimensionality of the feature space to a minimum. Feature selection is a very important computational preprocessing step for most subsequent research, such as building accurate inferential or classification models of the observed phenomenon, or elaborating the found dependencies (i.e. connections of the selected features to the target functions), perhaps even by means of laboratory/experimental studies.

However, feature selection is not an easy task, especially when the data is of high dimensionality and the relations are multivariate and nonlinear, and when there are many factors that weakly contribute to the target functions. Thus, brute force techniques are infeasible for such high dimensional problems due to its exponential time complexity and obviously well developed linear correlation methods to choose relevant variables for the learning task fail. In these cases, important variables may have even lower linear correlations with the target functions than the irrelevant variables. An alternative to complex learning machinery to catch such dependencies, such as neural networks, support vector machines, and other wrappers (Hsu and Lin 2002; Scholkopf and Smola 2002; Burges 1998; Bishop 1995; Vapnik 1995), a well known information theoretic measure of dependence, Shannon’s mutual information (Shannon 1948), works well for both linear and nonlinear relations. Mutual Information, abbreviated as

MI in text and in formulas as *I*(.;.), has recently been used for feature selection as a filter (sorting the variables from most relevant to the least) in several studies (Ding and Peng 2005; Peng, Long and Ding 2005; Kwak and Choi 2002). However, *MI* works well for the well sampled datasets (Endres and Foldiak 2005). Thus, in many areas of experimental sciences, it is a difficult task to calculate *MI* accurately due to the limited sample size. The work of Endres and Foldiak (2005) aims at alleviating this problem directly by reducing the number of bins of the input variables, which would increase the sampling rate of the joint sample space. Moreover, *MI* measure is not very sensitive to rare but predictive relations as is, because such rare relations may have little information content from information theoretic perspective. Scientists, on the other hand, may be interested in such minute relations. For example, in the field of biomedical, it would be very important to detect such relations, such as whether a variable is predictive about a rare type of cancer with few data samples in the experimental database. Such a variable should not be labeled unimportant just because there are a very large number of samples that belong to other phenotypes (many healthy subjects and many subjects with other cancer types in the study).

In this thesis, it will be showed that using *MI* can lead to missing features that can be very predictive of rare but suspicious coincidences. Instead, a novel measure of dependence based on *MI* that uses the concept of suspicious coincidences (Becker 1996) to fine tune the mutual information measure is proposed. Proposed measure works as a filter, which weighs more the samples with predictive powers, thus effectively eliminates the samples with no predictive contribution. This modification catches the suspicious coincidences well and also solves the low sampling rate problem mentioned above in a rather indirect way.

Moreover, these relations can be explored and experimented for deeper understanding and further insight into the field. If the aim of choosing the relevant variables was only to achieve the best possible prediction accuracy of a target variable disregarding the significance of rare classes, or without regarding the needs of scientific research, then *MI* would suffice. However, another important task is to guide the scientists to find out all (even small) relations among variables (Favorov and Ryder 2004; Kursun and Favorov 2004; Mjolsness and DeCoste 2001). Therefore, there are many research efforts on developing various measures of dependence between random variables. To mention a few, Gebelein’s maximal correlation, Rényi’s entropy, quadratic mutual information, Kernel mutual information, Kernel generalized variance, Kernel covariance, and so on are widely studied subject in statistics and used in many fields of scientific research (Shawe-Taylor and Cristianini 2004; Scholkopf and Smola 2002; Joe 1989; Breiman and Freidman 1985). Several types of mutual information have also been studied and various methods have been proposed to improve its usability in small datasets (Grande, Rosario, and Suarez 2008; Novoviccová et al. 2008; Baofeng and Nixon 2007; Valenzuela et al. 2006; Endres and Foldiak 2005; Peng, Long and Ding 2005; Kwak and Choi 2002; Becker 1996). In this thesis, we offer a novel approach based on conditional probabilities for adapting *MI* to such cumbersome datasets that have been faced in many fields of modern science.

The remaining of this thesis is organized as follows: Section 2 reviews wrapper, embedded, and filter methods which are most commonly used techniques in feature selection. In Section 2, Shannon’s entropy, mutual information, and the two most commonly used classifiers in the implementation of wrappers, artificial neural networks (ANNs) and support vector machines (SVMs) are reviewed. In Section 3, a demonstrative example is given to show the insensitivity of *MI* to rare events/classes and the derivation of our proposed measure called predictive mutual information (*PMI*) is presented with toy examples that visualize the accomplishment of *PMI*. Moreover, *PMI* is combined with mRMR (maximum Relevance Minimum Redundancy) approach by Peng et al. (2005) due to recognizing that the combinations of individually good variables do not necessarily lead to good classification/prediction performance. In Section 4, as a comparison, *MI* and *PMI* are used to identify the most relevant (predictive of the class labels) features of a real dataset, called Arrhythmia (Guvenir et al. 1997), and some of these features selected by *PMI* and *MI* are qualitatively compared by visualizing their joint distributions with the class labels. As part of the comparisons, a permutation test is applied on the dataset to judge the robustness of both measures, and finally, an SVM classifier is trained for a quantitative assessment of the joint information content of the best ranking features by these measures. In section 4, we also include some of the results of our analysis on sulfur dioxide (SO_2) dataset

(Identifying effective variables and building predictive models using mutual information and support vector machines for SO_2 concentration prediction, submitted article) which also shows that *PMI* can help detecting important features (according to the specialists in Environmental engineering) that can be missed by *MI*.

2. FEATURE SELECTION METHODS

One of the major problems in machine learning and related fields is feature selection. The main reason for feature selection's importance is that, in general, there are a large number of variables to choose from in order to be used in the learning/training process because using all the variables would simply worsen the generalization of the learning algorithm applied due to the phenomenon known as curse of dimensionality. The functions/relations to learn are generally multivariate and nonlinear, which prevents the use of simple techniques that would work without feature selection, such as mixture of experts, PCA, correlations, or linear discriminants.

Traditionally, the methods for feature selection are broadly divided into three categories: wrapper, embedded, and filter methods (Zhang and Deng 2007).

2.1 WRAPPER METHODS

The wrapper methods utilize the classifiers such as Support Vector Machines (SVM), Artificial Neural Networks (ANNs), etc., as evaluation functions and search for the optimal feature subset for the learning task, thus taking into account the joint effects of the variables.

A wrapper, in general, applies heuristic searches among exponentially many feature subsets, such as forward selection of features (starting from empty set of features and at each iteration adding the most "helpful" variable) or backward elimination of features (starting from all features and at each iteration removing the least "helpful" variable). In other words, for example, in the forward selection approach, firstly, the feature that has the best individual performance, i.e., the feature that has the best prediction accuracy over the target feature is chosen and then all possible combinations of that feature with the others are tested. Algorithm continues until sufficiently many features are selected or the classifier accuracy is high enough. However, for feature ranking, the algorithm continues until all combinations are tried and the computational complexity becomes

$O(n^2)$, where n is the number of features (which means that classification algorithm will be performed n^2 times). Similarly, backward elimination, nearly works in the same manner with forward selection. With this method, the first model is built by taking all the features into account. Then all the surviving features are removed one by one (with substitution) and the least useful feature identified is dropped in each iteration of this process. Obviously, this algorithm has $O(n^2)$ complexity, too.

Since wrapper methods need a classifier to be applied, the optimal values of chosen classifier's parameters that fit the data must be determined (which may vary from subset to subset, an additional complexity). Moreover, performing many trials this way may cause over-fitting problem (Reunanen 2003; Caruana, Lawrance, and Giles 2000) thus reduce the generalization of the classifier because basically we would be trying to choose the best subset that maximizes the prediction accuracy on our test set but the prediction accuracy on validation set would be compromised. Besides, especially with the use of under sampled data, dividing the data into subsets to use in the classification algorithm's training, testing, and validation steps may cause losing some important information, thus changing the order of features in feature ranking.

Redundancy of features that must be taken into account is another problem if the aim is feature ranking, i.e., if not selecting a compact and discriminative subset of features. When some of the features carry the same information about target feature, removing one of them does not affect the prediction accuracy, and so that variable seems irrelevant with the target feature. For example, in forward selection approach, if one of the redundant variables is already selected, then inclusion of the other redundant variables will not improve the prediction accuracy as if it has no relevance to the target. This will affect the ranking of features. So, if the aim is identifying the relevant features, for subsequent research, such as analyzing the found dependencies or building accurate and robust (if one variable is not available, using alternative variables) inferential models of the observed phenomenon, then redundancies prohibit the use of wrappers.

Two of the classifiers which are commonly used in the implementation of wrapper methods are described below:

2.1.1 Artificial Neural Networks

A neural network is basically a processing device implemented as an algorithm that takes the form of a network of many simple processing elements. Neural networks have a system by which the weights of the connections between the processing elements can be adjusted on the basis of patterns in a presented dataset. These weights can be adjusted, changing the initial state of network, so the system appears to 'learn'. The statistical potential for neural networks lies in their ability to generalize or even predict (Warner and Misra 1996).

Most commonly used feed-forward ANN is the multilayer perceptron (MLP) consisting of an input layer, an output layer and at least one hidden layer, making a total of at least three layers (Figure 2.1). On each layer there can be different number of nodes or neurons. The training of an MLP is based on back propagation error correction, which uses gradient descent optimization for error reduction. The training can be carried out either instantly or in batches. Instant training means the weights are adjusted instantly respective to the error of a batch of input patterns. Besides, to smooth out the training process, a learning rate and a momentum factor are often adapted in error correction (Jiang et al. 2004). A brief description of ANN's implementation is given below:

The activity of a hidden unit h is computed as a sigmoid function of the activities of its input sources:

$$H_h = \tanh(\sum_i w_{i,h} \cdot x_i)\,, \qquad (2.1)$$

where x_i is the value of input variable i, and $w_{i,h}$ is the weight of its connection onto the hidden unit h. The activity of the output unit y is:

$$y = \sum_{h} w_h \cdot H_h , \tag{2.2}$$

where w_h is the weight of the connection from the hidden unit h to the output unit.

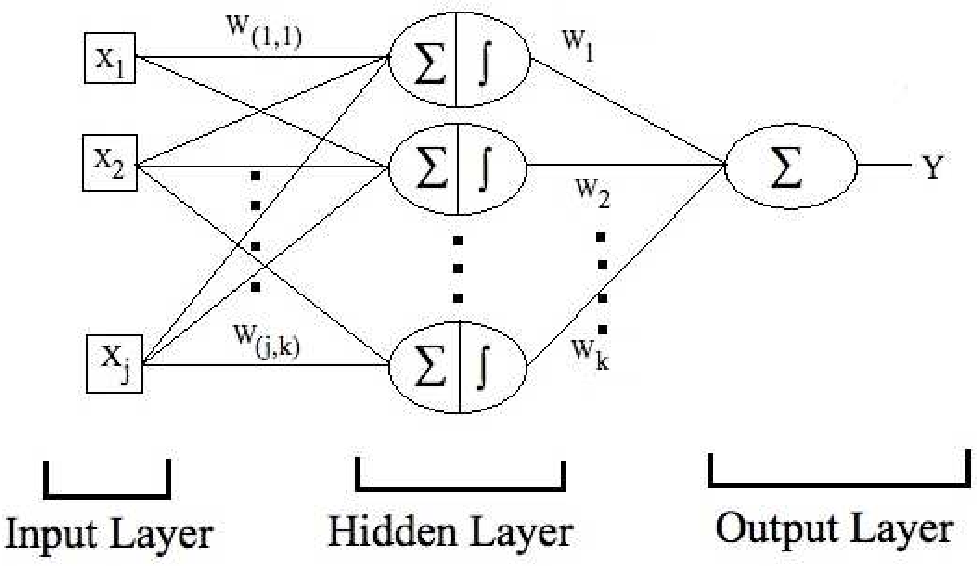

Figure 2.1 : An MLP structure with one hidden layer

The training signal T (the expected value of the target variable) is used to adjust the weights of connections, generally, by the well-known error backpropagation algorithm of Rumelhart, Hinton, and Williams (1986). Specifically, the error signal, δ is first computed as:

$$\delta = T - O. \tag{2.3}$$

For the hidden units, δ is backpropagated as:

$$\delta_h = \delta \cdot w_h \cdot (1 - H_h^2). \tag{2.4}$$

Connection weights are adjusted by:

$$\Delta w_{i,h} = \mu_i \cdot I_i \cdot \delta_h \text{ and } \Delta w_h = \mu_h \cdot H_h \cdot \delta \,, \tag{2.5}$$

where μ_i and μ_h are learning rate parameters for the input and hidden unit connections, respectively.

Determining the optimal number of layers and the number of neurons is a process of trial and error. Wang et al. (1984) suggest that it is the complexity of the dataset that controls input and output neuron numbers, and although several empirical rules have been suggested it is likely that the hidden neuron number is problem-specific (Spellman 1999). Since the input and output layers have fixed number of neurons, in practice, the best appropriate model performance is found by adjusting the number of hidden layers (and hidden units in each layer).

2.1.2 Support Vector Machines

SVM is a more modern classifier that uses kernels to construct linear classification boundaries in higher dimensional spaces and they generalize very well. Here a brief intuitive explanation of the SVM approach is provided using a geometric perspective. To visualize how SVMs work, a graphic example of hypothetical data described by two variables, X1 and X2, which is shown in Figure 2.2 will be used, with data samples divided into two classes. The described concepts, however, are generalizable to larger numbers of input variables and also to regression tasks (i.e., learning a continuous function of the input, rather than its class partitions).

The fundamental SVM design is built on several key insights (Vapnik 1995; Vapnik 1998; Schölkopf 2002). The first insight is to use the “optimally” placed decision hyperplane to separate the sample classes. For example, in Figure 2.2A the training data samples belonging to two different classes cluster separately in different regions of the input space (i.e., the space defined by input variables X1 and X2) and can be easily separated by a line. This line can be used to classify new, test data samples, according to their position relative to this line. In Figure 2.2A, two among many possible placements of the decision line are shown. While they separate the two groups of training samples equally well, more preferable is the black line. This line is placed so as to maximize the minimal distance between it and the training samples, and it is more preferable because it is less likely to make false classification decisions on future samples.

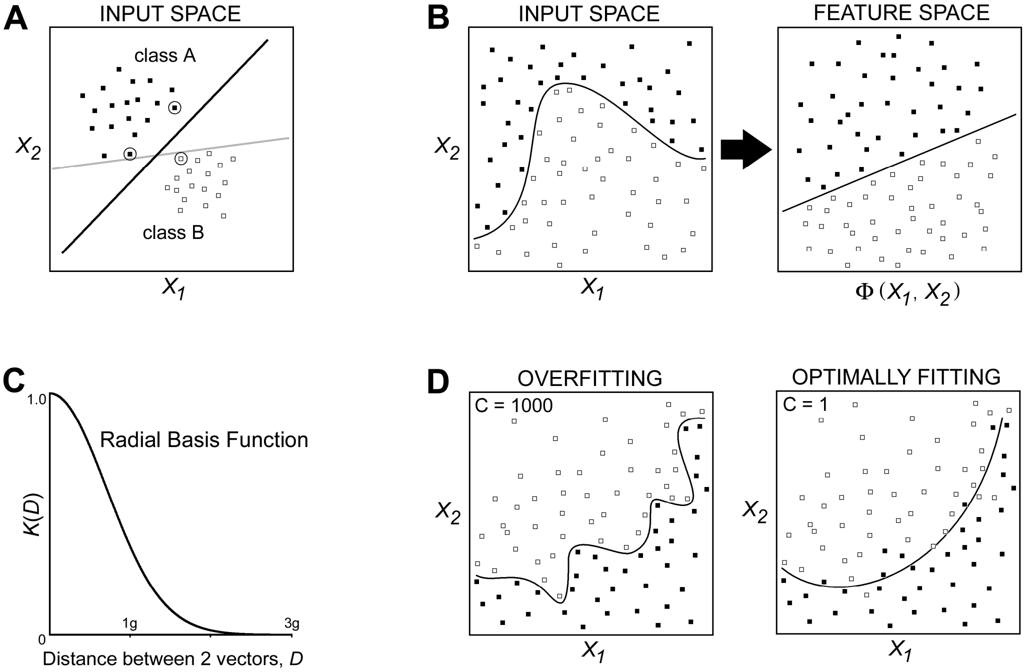

Figure 2.2 : Key features of SVM design. A: Optimal decision hyperplane. Little black squares are data samples of class A, little open squares are data samples of class B. The optimal boundary between the two classes is shown as the black line. Circled samples are the "support vectors," which determine the orientation of the optimal boundary. B: Transformation of Input space into Feature space. The Feature space has more dimensions than the Input space, but only two are shown in this illustration for display clarity. C: Radial Basis Function. The value of the function is plotted against the distance between two vectors, which is expressed as a fraction of the g-parameter. D: Control of the smoothness of the class boundary. The data samples are shown in the input space and are separated into two classes by a curved line. In the left panel, the highly-convoluted boundary is overfitting: it correctly separates all the shown data samples by their classes,

but is likely to be less accurate on new data samples than the smoother boundary in the right panel

Note that the placement of the optimal decision hyperplane (the line in Figure 2.2A) is determined not by all the training samples, but only by the samples closest to the hyperplane (they are indicated in Figure 2.2A by circles). Such training samples that determine the orientation of the decision hyperplane are called the “support vectors.” Use of the optimal decision hyperplane is the foundation of the SVM superior ability to generalize from training samples to new data.

The second insight is to make classification (or regression) decisions not in the input space (defined by the input variables), but in a “feature” space. This distinction becomes important when the training data samples cannot be separated in the input space by a hyperplane. For example, in Figure 2.2B the two classes cannot be separated completely by a straight line, but only by a curved line. Unfortunately, finding the optimal curved partition of the input space is much more difficult. Unlike finding optimal linear partitions, finding optimal nonlinear partitions takes much longer time and is quite likely to produce suboptimal solutions (become trapped in local minima). We can overcome this problem; however, if we would somehow transform the input space into such a new “feature” space, in which the sample classes become linearly separable (see Figure 2.2B). Then we can use the techniques of linear separation on the transformed data and determine their optimal partition in the feature space.

The third insight is that explicit remapping of the data from the input space to a feature space does not have to be actually done in practice. Evaluating data points in a feature space can be replaced, with exactly the same results, by simply evaluating data points in the original input space using an appropriate kernel function. A very popular kernel function is the Radial Basis Function (RBF). It expresses similarity of two vectors, and, as a function of the Euclidean distance between them, D_{ij}, according to An RBF kernel is shown in Figure 2.2C. The RBF parameter g controls the width of the kernel.

RBF kernel has been found to be very effective in a wide range of SVM applications. In principle, for problems of a particular nature, there might be a special kernel that will be most effective in separating different sample classes there. However, finding such an optimal problem-specific kernel usually is not practical, and use of a known, “general-purpose” kernel (such as RBF) will still provide a reasonably successful solution. RBF is generally the first kernel type to try; if it fails, other common kernels provided by software packages (in particular, polynomial and sigmoid kernels) can be tried next.

The fourth insight concerns the danger of SVM overfitting on the training data. As illustrated in Figure 2.2D, two sample classes might have partially overlapping distributions in the input space. Using kernels, we will be able to achieve 100% separation of samples belonging to the two classes by fitting a highly convoluted boundary to them (see the left panel in Figure 2.2D). But this boundary will be mistaken, being misled by noise in the data. A much less convoluted boundary (as shown in the right panel in Figure 2.2D) will be, objectively, more accurate, reflecting the true interface between the two class distributions. Thus, by setting limits on the degree of acceptable complexity of the SVM-drawn boundaries, we might be able to improve the SVM performance on future, test data samples, despite doing worse on the training data. An SVM parameter that controls the complexity of class partitions is known as “penalty error,” or parameter C. As C decreases in value, the boundaries become smoother. As a rule, when fewer numbers of data samples are available for training an SVM, the attempted class partitions should be more constrained in their complexity by reducing the value of C-parameter. Parameter C enhances SVM ability to generalize successfully from training samples to new data.

In conclusion, in order to use an SVM on a particular dataset, only three basic parameters have to be specified: (1) C-parameter; (2) the choice of the kernel (RBF is recommended first); and (3) a kernel-specific parameter (e.g., g-parameter for RBF, or degree of the polynomial for polynomial kernel). The optimal values of these parameters are problem-specific and are determined empirically by trial-and-error procedures.

2.2 EMBEDDED METHODS

Embedded methods perform selection of features during the training process of the classifier such as weight decay in neural networks (Bishop 1995). They are more efficient than wrappers in several aspects: they make better use of the available data by not needing to split the training data into a training, testing, and validation set; they reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated (Guyon and Elisseeff 2003). However, just like wrapper methods, embedded methods are specific to the particular learning algorithm.

As an exemplary formulation we will describe how weight decay works in neural networks. Weight decay adds a penalty term to the error function, usually sum of squared weights times a decay constant, γ, which reflects to the update rules as:

$$\Delta w_{i,h} = \mu_i \cdot I_i \cdot \delta_h - \gamma \cdot w_{i,h} \text{ and } \Delta w_h = \mu_h \cdot H_h \cdot \delta - \gamma \cdot w_h \,. \qquad (2.6)$$

The weights are forced towards zero by reducing them relative to their strengths with a decay parameter between 0 and 1. The input variables with nearly zero weights, then, can be assumed to be irrelevant to the learning task.

2.3 FILTER METHODS

A filter selects features without involving any classifier/regressor for evaluation and it is based on a measure of relevance/dependence to the target such as the two frequently used measures, Pearson correlation, and mutual information.

2.3.1 Pearson Correlation Coefficient

Correlation coefficient, ρ, between two signals x_i and y_i is a well-known measure of how highly two signals correlate, which is computed as follows:

$$\rho = \frac{N \cdot \sum_i (x_i \cdot y_i) - \left(\sum_i x_i \cdot \sum_i y_i \right)}{\sqrt{\left(N \cdot \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right) \cdot \left(N \cdot \sum_i y_i^2 - \left(\sum_i y_i \right)^2 \right)}} \tag{2.7}$$

where *N* stands for the number of observations.

While Pearson's correlation coefficient (Stigler 1968) *corr(X;Y)* is the basic tool to describe a degree of dependence between two random variables, it is a linear measure and obviously the equality *corr(X;Y)=0* does not imply independence of *X* from *Y*. However, in real datasets, the functions/relations to learn are generally multivariate and nonlinear, and in these cases important variables may have even lower linear correlations with the target functions than the irrelevant variables.

2.3.2 Mutual Information

Mutual Information is a classical and widely used measure of dependence that serves as a good feature ranking and selection algorithm. *MI* has recently been used for feature selection and ranking as a filter (sorting the variables from most relevant to the least) in several studies in many fields - medicine, neuroscience, genomics and related fields, ecology, economics, etc (Ding and Peng 2005; Kwak and Choi 2002; Peng, Long, and Ding 2005).

Shannon’s entropy (Shannon 1948) is a measure of the uncertainty of a random variable *X* and thus, it quantifies how difficult to predict that variable. The entropy of a random variable *X*, denoted *H(X)*, is a functional of the probability distribution function *P(X)*, and is sometimes written as *H(P(X))*. Because, the entropy of *X* does not depend on the actual values of *X*, it only depends on *P(X)*.

The definition of Shannon's entropy can be written as an expectation:

$$H(X) = -E[\log P(X)]| = -\sum_{x}\left[p(x)\log(p(x))\right], \qquad (2.8)$$

where *p(x)* = *P(X=x)* is the probability distribution function (more it is the precisely probability mass function for the discrete case but the results are generalizable) of *X*. Hence the Shannon's entropy is the average amount of information contained in random variable *X*. In other words, it is the uncertainty removed after the actual outcome of *X* is revealed.

Mutual Information (abbreviated as *MI* in text and in formulas as *I*(.;.)) is a measure of mutual dependence of the two variables based on the entropy:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \qquad (2.9)$$

or similarly,

$$I(X;Y) = H(P(X)) + H(P(Y)) - H(P(X,Y)). \qquad (2.10)$$

MI can be conceptually visualized in its relation to the entropies of the two variables and their common information (certainty about the state of one variable by using the state of the other) as in Figure 2.3 (MacKay 2003):

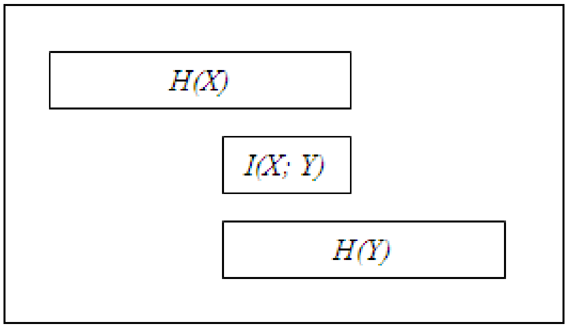

Figure 2.3 : Visualization of mutual information

Shown as the intersection of *H(X)* and *H(Y)* is the amount of information (mutual information) that can be predicted about *Y* knowing the values of *X*. Normalizing this value can be more interpretable to use for feature ranking because it makes more intuitive sense. It can be defined as:

$$N(X;Y) = \frac{I(X;Y)}{H(Y)} \tag{2.11}$$

where; *N(X;Y),* how much of the uncertainty (per cent) in *Y* is removed by knowing the present value of the feature *X*.

The measure *I* is also the KL divergence of the product *P(X)P(Y)* of the two marginal probability distributions from the joint probability distribution, *P(X,Y)*.

$$I(X;Y) = D_{KL}(P(X,Y) \parallel P(X) \cdot P(Y)) = \sum_{x}\sum_{y}\left[p(x,y)\log(\frac{p(x,y)}{p(x) \cdot p(y)}) \right] \tag{2.12}$$

where *p(x,y) = P(X=x, Y=y).*

In other words, *I(X; Y)* is the expected number of extra bits that must be transmitted to identify *X* and *Y* if they are coded using only their marginal distributions instead of the joint distribution.

2.3.3 Maximum Relevance – Minimum Redundancy

Maximum Relevance – Minimum Redundancy (mRMR) approach based on *MI* by Peng et al. (2005) aims to maximize the joint dependency of the selected variables by reducing the redundancies among them due to recognizing that the combinations of individually good variables do not necessarily lead to good classification performance. In other words, mRMR suggests incrementally selecting the maximally relevant variables while avoiding the redundant ones with the aim of selecting a minimal subset of variables that represents the problem. This helps (not guaranteed) the top m features selected most likely has the highest joint dependency.

According to mRMR approach, mth feature chosen for inclusion in the set of selected variables, *S*, must satisfy the below condition:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right], \qquad (2.13)$$

where *X* is the whole set of features; c is the target variable; x_i is the i^{th} feature.

In other words, the feature that has the maximum difference between its mutual information with target variable and the average mutual information with the features in S will be chosen next.

3. PROPOSED METHOD

3.1 PROPOSED METHOD: PREDICTIVE MUTUAL INFORMATION

Due to its superiority over linear methods such as Pearson's correlation coefficient, *MI* is a suitable technique for feature selection and ranking. However, *MI* requires a well sampled database, which is not typical of many fields of modern science (such as biomedical), in which there are limited number of samples to learn from, or at least, not all the classes of the target function (such as certain phenotypes in biomedical) are well-sampled. Moreover in such settings, each feature, among many, contributes in small amounts to the target function to be predicted, analyzed, or modeled.

A demonstrative example for showing this problem with *MI* is presented below:

Figure 3.1 shows the joint distribution of 100 data points with two attributes, *X* and *Y*, where each one takes discrete values from 1 to 10 (for display purposes, to be able to show the density of the points at (*X*=i, *Y*=j), small random noise is added to the data rather than making it a 3D plot for the PDF).

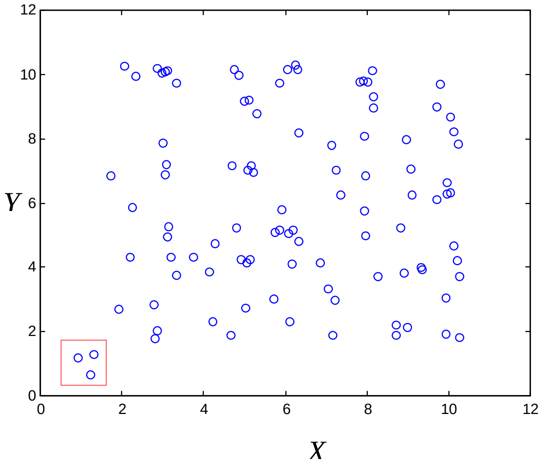

Figure 3.1 : Joint distribution of two dependent variables (with the most prominent relation shown in the red square) with 100 samples where *I(X; Y)* = 0.7194

It is a small information gain but could be very important to realize that we have "suspicious coincidences" enclosed in a red square in Figure 3.1 in the lower left corner where *X*=1, *Y*=1. There are 3 points/samples in this corner. 97 out of 100 samples, almost no orderly relations are noticeably present in the joint distribution. Therefore, naturally, *X* and *Y* have very low mutual information of 0.7194. What does this number 0.7194 mean to anybody, especially to a practitioner, say, in the biomedical field? Not much. The mutual information of *X* and *Y* is actually tiny; in fact, to have a better sense that it is really small, the mutual information is calculated of *X* and *Y* in the data shown in Figure 3.2, which is, this time, completely independently randomly distributed. Surprisingly, it is reported that these two random variables, which are generated so as to have no obvious relationship to each other (joint distribution given in Figure 3.2), has more mutual information than the dependent variables *X*, *Y* given in Figure 3.1.

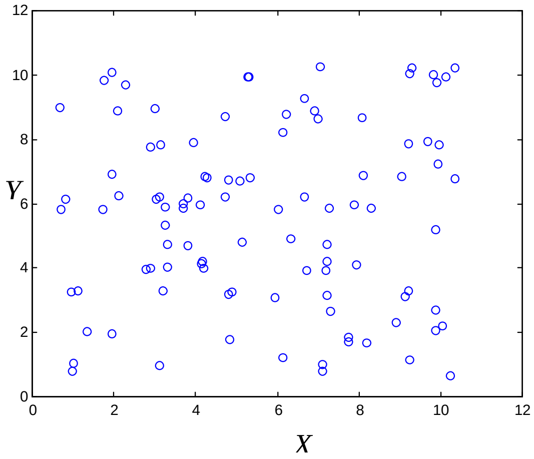

Figure 3.2 : Joint distribution of two independent variables with 100 samples where *MI*(*X*; *Y*) = 0.7993

Consider, *X* is a particular test result and *Y*=1 is an indicative of the fact that the patient is developing a certain type of cancer. In that case, this would be an important fact to realize that *X*=1 implies *Y*=1 (and vice versa), which has been seen 3 out of 100 observations (patients). In fact, if a big event, such as an earthquake happened 3 times when it was full-moon (or very hot weather), we would be very interested about these relationships (even if the experts claim otherwise). We must be extremely cautious evaluating data about rare classes (events). This situation is a typical one in the biomedical field; just because 97% is healthy (or with other types of disorders), it would be inappropriate to discredit the importance of rare but important classes such as certain disorders in the overall domain.

Note also that in the continuous-case (consider the dots in Figure 3.1 as actual sample points rather than having random scatter added for better display of the distribution), things get even worse because of the issues in probability distribution function (PDF) estimation using kernels or Parzen windows (Bishop 1995); or similarly, if we decide to discretize the continuous variables *X* and *Y* in order to calculate entropy and mutual information easier but at the cost of losing precision and accuracy. For example, using an equal width discretization as shown in Figure 3.3 results in a terrible discretization for this particular example that reduces the importance of the suspicious coincidences at *X=1, Y=1*. Equal frequency discretization or K-means clustering (Bishop 1995) would, leave alone stressing them, smear them with the other samples falling into the same bin. Moreover, adjusting the parameters of such methods can be troublesome.

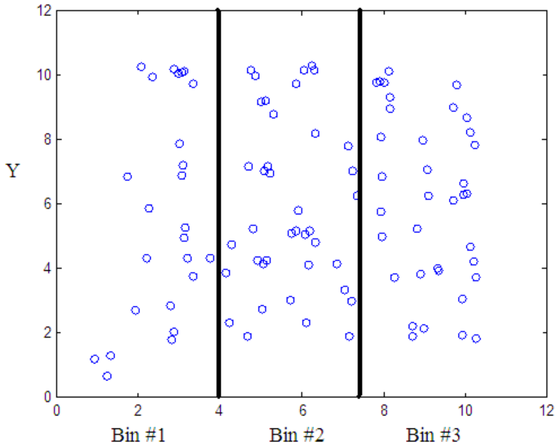

Figure 3.3 : Equi-width discretization worsens the importance of the *X*=1⇔*Y*=1 regularity

Yet, the situation gets worse for *MI*, when (*X*=1, *Y*=1) becomes more rare (when the probability of sample points increase in the space $X \geq 2$ and $Y \geq 2$). From Eq. 2.9, it can be related to the fact that mutual information between *X* and *Y* is small because the rare event at (*X*=1, *Y*=1) will add tiny amounts to all of the entropies *H(X), H(Y),* and *H(X,Y).* Thus, the mutual information will be very low, because what entropy measures is the average information content per sample. Entropy is an additive measure of information and it is proportional to the uncertainty of a random variable. In order to detect the predictability relations more precisely, it is needed to avoid the average-out effect of entropy on rare but well-predictable classes that we call suspicious coincidences (inspired by Becker 1996; Favorov and Ryder 2004; Kursun and Favorov 2004) by weighing them higher in the summation in Eq. 2.8.

Of course, conceptually, if some event is too rare it can be considered to be an outlier and ignored; but that is only if that event is not inferentially useful. In our toy example, even though the orderliness of the 97% of the joint distribution is negligible, the three percent of it is very orderly in predicting $Y=1$. Therefore, these relations must be amplified rather than blurred out. After all, the task must be finding orderly relations and automating the use of such orderly relations for building inferential models as in Virtual Scientist (Kursun and Favorov 2004). As a matter of fact, on the contrary to the engineering perception of many papers published on predicting the states of certain disorders from the input variables, such predictions are undoubtedly not the only goal of scientific research in the biomedical field.

To fine tune the mutual information aiming to use the concept of these suspicious coincidences (Becker 1996), we propose a novel measure of dependence, Predictive Mutual Information (*PMI*), that is based on *MI*.

3.1.1 Derivation of the Proposed Measure *PMI*

Realizing that having no samples in certain parts of the sample space is also valuable information because it increases conditional probabilities elsewhere, thus a form of mutual information is formulated as described below:

$$Q'(X,Y) = P(X,Y) \cdot P(X \mid Y) \cdot P(Y \mid X). \tag{3.1}$$

Q' is a modified joint probability distribution of X and Y, it basically weighs each p(x,y) entry by p(x | y) * p(y | x). This weight is between 0 and 1 and nonzero when p(x,y) is nonzero. Therefore,

$$0 \leq Q'(X,Y) \leq P(X,Y), \tag{3.2}$$

which must be followed by a normalization step as follows:

$$Q(X,Y) = \frac{Q'(X,Y)}{\iint\limits_{X,Y} Q'(X,Y)} \tag{3.3}$$

such that

$$\iint\limits_{X,Y} Q(X,Y) = 1.0\,. \tag{3.4}$$

Q' is normalized to Q so that it sums up to one as a well-formed probability distribution function, where *Q(X)* and *Q(Y)* are the marginal probability distributions obtained from *Q(X,Y)* as in Eq. 3.5 and 3.6, respectively:

$$Q(X) = \int\limits_{Y} Q(X,Y)\,, \tag{3.5}$$

$$Q(Y) = \int\limits_{X} Q(X,Y)\,. \tag{3.6}$$

PMI should be defined as such to avoid the average-out effect of *MI* on the rare but predictable classes and give some precedence to the predictable relations between the variables. Thus, *PMI* is defined as follows:

$$PMI(X;Y) = PMI(P(X,Y)) = I(Q(X,Y)) = H(Q(X)) + H(Q(Y)) - H(Q(X,Y)), \tag{3.7}$$

Moreover, *Q'* can be written in its general form as:

$$Q'(X,Y) = P(X,Y) \cdot P(X \mid Y)^{\alpha} \cdot P(Y \mid X)^{\beta}. \tag{3.8}$$

In this formulation, *Q'* serves as a filter on the PDF of the data, *P*, and passes the important probabilities in *P* based on α and β. Therefore, we can control the sort of entropy to include in the naïve mutual information calculations using this filter with various nonnegative values of α and β.

Also note that:

$$Q'(X,Y)= \frac{P(X,Y)^{\alpha+\beta+1}}{P(X)^{\beta}\cdot P(Y)^{\alpha}}. \tag{3.9}$$

Conceptually, *Q'*, or its normalized version *Q*(*X*,*Y*), gives us a measure of predictability of *X* and *Y* from each other proportional to their joint frequency. Clearly, when $\alpha=\beta=0$, we have *PMI* equal to Shannon's mutual information *MI*. Although it is out of the scope of this paper, briefly we can state that, when $\alpha=\beta=1$, it relates to SINBAD of Favorov and Ryder (2004), IMAX of Becker (1996), ACE algorithm of Breiman and Freidman (1985). When $\alpha=1$ and $\beta=0$, or $\alpha=0$ and $\beta=1$, we obtain a sort of Bayesian. When $\alpha=\beta=1/2$, we obtain links to Rényi generalized divergence of order 1/2 (or equivalently a sort of the Bhattacharyya coefficient).

Clearly, using the notation $q(x) = q(X=x)$, $q(y) = q(Y=y)$, and $q(x,y) = q(X=x, Y=y)$, Eq. 3.11 above can be written as:

$$PMI(X;Y) = -\sum_{x}\left[q(x)\log(q(x))\right] - \sum_{y}\left[q(y)\log(q(y))\right]$$
$$+\sum_{x}\sum_{y}\left[q(x,y)\log(q(x,y))\right], \tag{3.10}$$

$$PMI(X;Y) = D_{KL}(Q(X,Y)\,\|\,Q(X)\cdot Q(Y)) = \sum_{x}\sum_{y}\left[q(x,y)\log(\frac{q(x,y)}{q(x)q(y)})\right]. \tag{3.11}$$

3.1.2 What Does *PMI* Accomplish (Example Revisited)

PMI measure gives 1.5031 versus 1.1986 for the two datasets given in Figures 3.1 and 3.2, respectively. Figure 3.4 below shows what is accomplished by the *PMI* measure. In the right panel of Figure 3.4, *Q* works as a filter passing "interesting" regions (suspicious coincidences) of the distribution shown in the left panel (also in Figure 3.1). Among the most obvious of these suspicious coincidences are $X=1 \Leftrightarrow Y=1$ and also where *X*=5 it turns out with *Y*=6 and vice versa. Thus, it simplifies the PDF and allows mutual information calculation to reflect predictability of variables from each other. *PMI* is 1.5031, which significantly surpasses the *PMI* of 1.1986 of the independent random variables given in Figure 3.2.

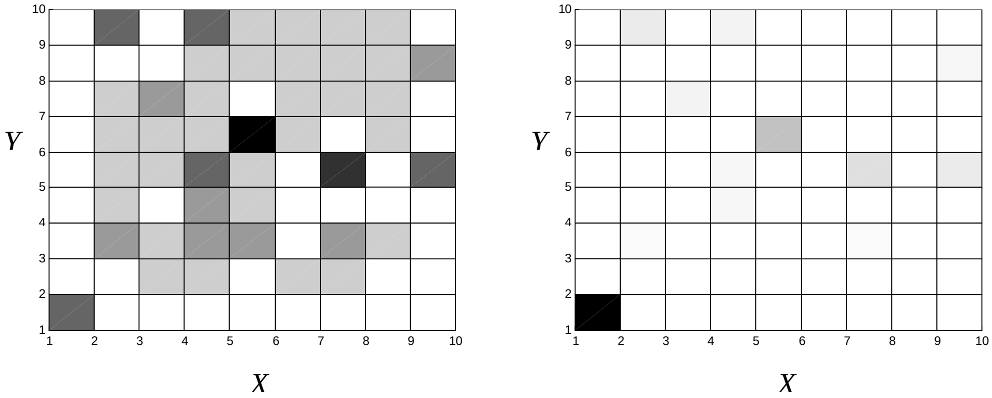

Figure 3.4 : (Left) A different view of Figure 3.1. The intensity of each cell is inversely proportional (darker for higher probability) to $P(X, Y)$; (Right) $Q(X,Y)$ with α=1 and β=1

3.1.3 Simulations for Showing the Statistical Significance of Power of *PMI*

The results presented in Section 3.1.2 belong only to a single run of the example and may or may not hold true for some other runs (i.e. different random selections of the 100 points in 2D plots in Figures 3.1 and 3.2 such that in former, there are exactly three points with X=1 which also their Y=1, and vice versa. The other points are uniformly randomly distributed between (2, 2) to (10, 10)). To determine to what extent the results are statistically significant, we have simulated the example 100 times with different random data points generated for both dependent (Figure 3.1 example with $X=1 \Leftrightarrow Y=1$) and independent (Figure 3.2 example) cases.

Figures 3.5 and 3.6 show, respectively, the distribution of *PMI* and *MI* measures for the 100 runs of the demonstrative example given in Sections 3.1 and 3.1.2 (a total of 300 data points are used in the experiments, instead of 100). The scores are shown in red for the dependent case and in blue for the independent case. Clearly, *PMI* for the dependent variables is higher than it is for the independent ones (i.e. blue bars are far to the left of the red ones). However, *MI* fails to provide discriminative scores (i.e. blue and red bars are intermixed).

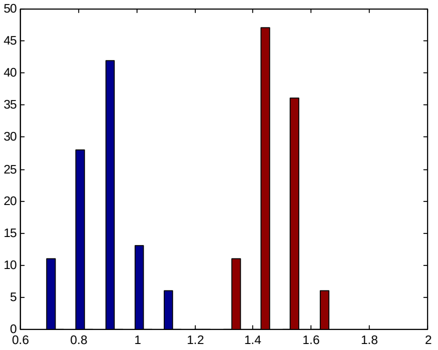

Figure 3.5 : Distribution of *PMI* measure for the demonstrative example

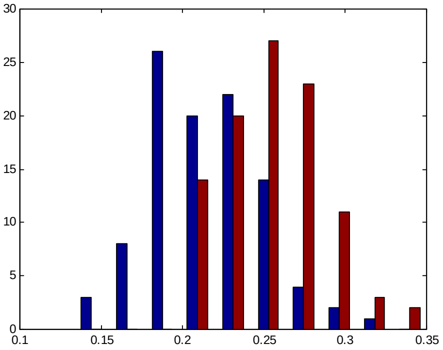

Figure 3.6 : Distribution of the Shannon's Mutual Information measure for the demonstrative example

3.1.4 More Demonstrative Examples

In this section, there are more demonstrative examples that show the behavior of *PMI*. In all the examples, $\alpha=\beta=1$ is used for simplicity (it makes sense to use a lower α as well because the features are eventually to be used to predict the target class). As *PMI* will be used for feature selection ultimately in section 3.2 (experimental studies and results), in this section, the terms *the feature* (or the feature value) and *the class label* (or class #) will be used. To relate to the terminology used in the previous sections, *the feature* refers to *X* and *the class label* refers to *Y* (these terms will be used interchangeably).

Similar to Figure 3.4, the left panels of the Figures 3.7, 3.8, 3.9, 3.10 show several interesting PDFs and the right panels show the Q-PDFs, *Q*(*X*,*Y*), as defined in Section 3.1.1. Also, just like Figure 3.4, the intensity of each cell in the PDFs in these figures is inversely proportional (darker for higher probability) to its probability.

Applying *PMI* to the feature, whose plot versus the class label is given in the left panel of Figure 3.7, the given PDF is converted into Q-PDF shown in the right panel of the same figure. *PMI* simply ignores the regions that have little mutual predictive information and emphasizes the "suspicious coincidences" of the feature with the class labels. Thus, the information the feature carried that was important but blurred due to its rarity is made clearer.

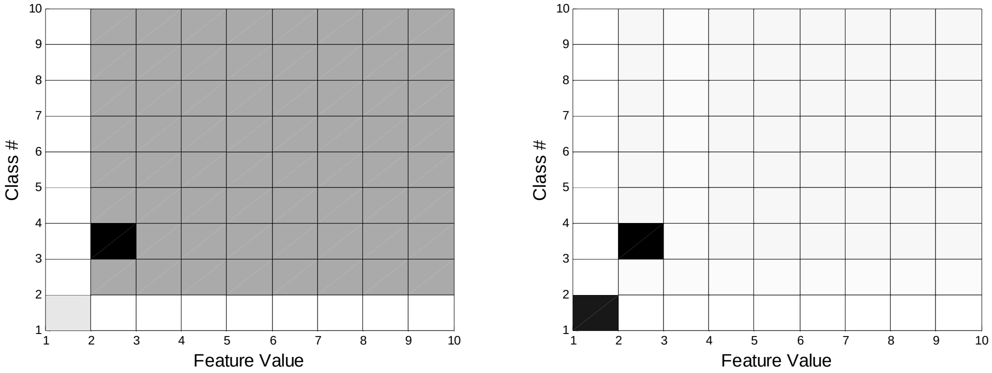

Figure 3.7 : (Left) Joint distribution of the feature and the class label (*MI* score is only 0.0501); (Right) after applying *PMI* filter, *MI* (or *PMI*) score is 0.9633

Another demonstrative example is shown in Figure 3.8. This time the feature seems completely irrelevant to the class labels according to *MI* measure since it has a very low *MI* score of 0.0133. However, after applying *PMI* filter, PDF becomes as shown in the right panel of Figure 3.8, which shows a clear relation of the feature with the target class.

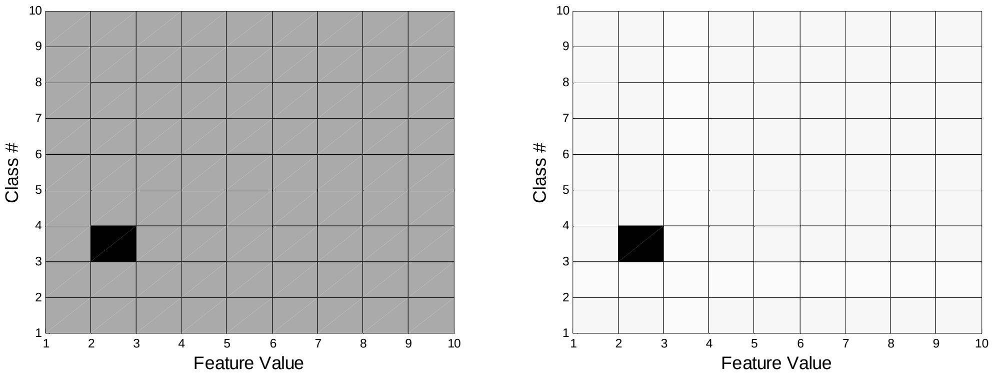

Figure 3.8 : (Left) Joint distribution of the feature and the class label (*MI* score is only 0.0133); (Right) after applying *PMI* filter, *MI* (or *PMI*) score is 0.3002

Figure 3.9 shows our next demonstrative example, which has an interesting scenario. Each class except class #1 has 30 samples for each feature value from 2 to 10, inclusive (i.e. a total of 270 samples for each of class from 2 to 10). There are only 10 samples that belong to class 1 and for all of these samples, the feature value is 1. The joint distribution is shown in the left panel of Figure 3.9. *MI* value of this variable is only 0.0384. It cannot be concluded that this variable contains predictive information about the target class using this *MI* score which is approximately zero. However, the right panel of Figure 3.9 shows the Q-PDF after applying *PMI*, for which *PMI* score is 0.8113. From this plot and the *PMI* score, one can easily conclude that when the feature value is 1, the sample must belong to class 1, which would be significant information, say, if class #1 refers to an important but rarely seen type of cancer.

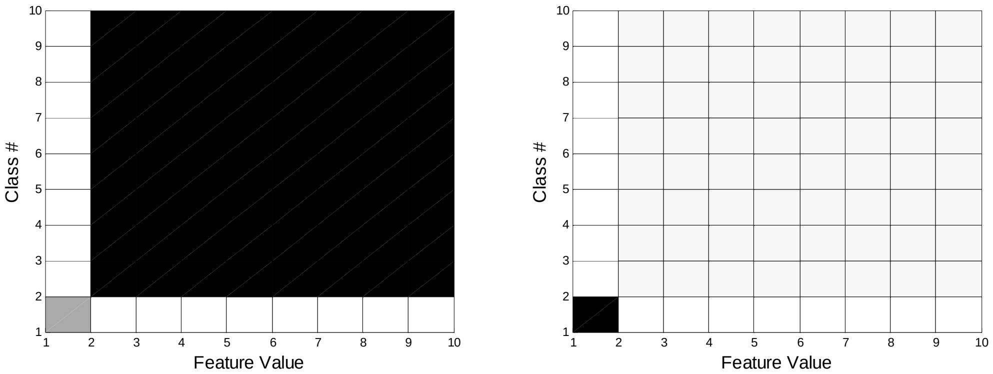

Figure 3.9 : (Left) Joint distribution of the feature and the class label (*MI* score is only 0.0384); (Right) after applying *PMI* filter, *MI* (or *PMI*) score is 0.8113

Last demonstrative example shown in Figure 3.10 looks like the example in Figure 3.9. The only difference between these two examples is that the relation *X=1* ⬄*Y=1* is more prominent here. In the example shown in Figure 3.10, each class except class #1 has 10 (instead of 30 for the example in Figure 3.9) samples for each feature value from 2 to 10, inclusive (i.e. a total of 90 samples for each of class from 2 to 10). There are only 10 samples that belong to class 1 and for all of these samples, the feature value is 1. *MI* score is low again, 0.0950. However, after applying *PMI* filter, the interesting regions of the PDF are emphasized, thus, a clearer figure that shows the relation of the variable with the target class emerges as shown in the right panel of Figure 3.10. *PMI* score of the variable is 1.0000 (not to confuse with the correlation coefficient 1.0, this is just a measure of mutual information), which is high enough to confirm that there is a relation between this variable and the target class.

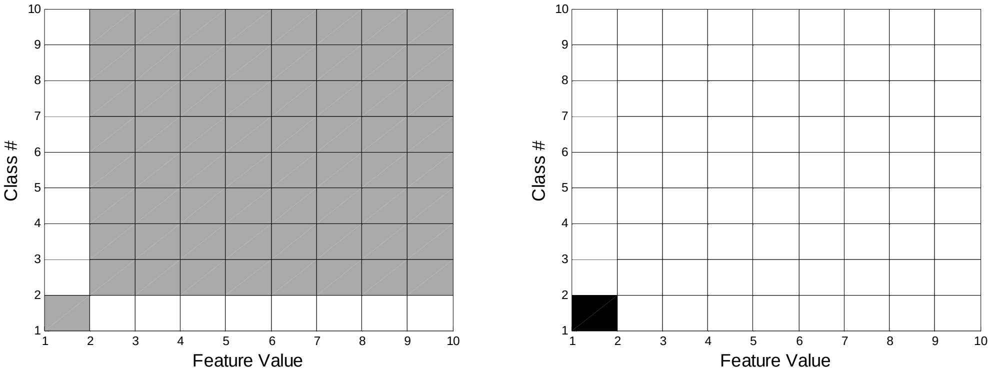

Figure 3.10 : (Left) Joint distribution of the feature and the class label (*MI* score is only 0.0950); (Right) after applying *PMI* filter, *MI* (or *PMI*) score is 1.0000 (not to confuse with the correlation coefficient 1.0, this is just a measure of mutual information)

3.2 EXTENSION OF *PMI* COMBINING WITH mRMR

Proposed measure *PMI* can be combined with mRMR approach by Peng et al. (2005) due to recognizing that the combinations of individually good variables do not necessarily lead to good classification/prediction performance. In other words, to maximize the joint dependency of top ranking variables on the target variable, the redundancy among them must be reduced, which suggests incrementally selecting the maximally relevant variables while avoiding the redundant ones. This helps (not guaranteed) the top m features selected most likely has the highest joint dependency.

According to mRMR approach, mth feature chosen for inclusion in the set of selected variables, *S*, must satisfy the below condition:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right], \tag{3.12}$$

where *X* is the whole set of features; *c* is the target variable; x_i is the i^{th} feature.

In other words, the feature that has the maximum difference between its mutual information with target variable and the average mutual information with the features in S will be chosen next.

For combining *PMI* with mRMR, straightforward approach is using *PMI* instead of *MI* in equation 3.12 as in the below equation:

$$\max_{x_j \in X - S_{m-1}} \left[PMI(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} PMI(x_j; x_i) \right]. \tag{3.13}$$

However, reason of suggesting *PMI* measure instead of *MI* is trying to catch the suspicious coincidences between the features and the target class (especially about under sampled classes). We are not interested in the minute relations among the features. Therefore, while calculating the redundancies between the candidate variable and the selected variables, using *PMI* instead of *MI* does not make sense. Considering this situation, Eq. 3.13 can be rewritten as below:

$$\max_{x_j \in X - S_{m-1}} \left[PMI(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]. \tag{3.14}$$

However, *PMI* and *MI* must be in the same scale to be used in the same equation. Let us go back and examine Eq. 3.12 if its terms are in the same scale. First term of the equation, $I(x_j, c)$, denotes the mutual information between the candidate variable and target class. Second term (can be called as redundancy term), $\frac{1}{m-1}\sum_{x_i \in S_{m-1}} I(x_j; x_i)$, measures the average redundancy between the candidate variable and selected the selected variables. These terms can be expressed in proportional to *H(x$_j$)* as shown below in Eq. 3.15:

$$\max_{x_j \in X - S_{m-1}} \left[H(x_j) \left(\frac{I(x_j, c)}{H(x_j)} - \frac{1}{(m-1)H(x_j)} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right) \right]. \tag{3.15}$$

In the above equation, first term, $\frac{I(x_j, c)}{H(x_j)}$ denotes how much entropy of the candidate variable x$_j$ in percent is common with the target class c. The second term $\frac{1}{(m-1)H(x_j)} \sum_{x_i \in S_{m-1}} I(x_j; x_i)$ measures how much entropy of the candidate variable x$_j$ in percent (average) is common with the selected variables. Multiplying the difference of these terms with the variable's entropy gives the unique information that the variable has about the target class. Among the candidate variables, the variable with the maximum value is chosen next for the selected set of variables. Obviously, this equation (Eq. 3.15) can be simplified to Eq. 3.12.

We are now ready to rewrite the Eq. 3.14 to calculate the difference of the terms in the same scale. Substituting $\frac{I(x_j, c)}{H(x_j)}$ with $\frac{PMI(x_j, c)}{H(Q(x_j))}$ in Eq. 3.15 gives the equation that combines *PMI* with mRMR. Finally, the equation that combines mRMR with the proposed *PMI* can be written as in Eq. 3.16:

$$\max_{x_j \in X - S_{m-1}} \left[H(x_j) \left(\frac{PMI(x_j, c)}{H(Q(x_j))} - \frac{1}{(m-1)H(x_j)} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right) \right]. \tag{3.16}$$

4. EXPERIMENTAL STUDIES AND RESULTS

4.1 ARRHYTHMIA

To illustrate the use of *PMI*, in this section, unless stated otherwise, α and β values are set to 1.0 empirically for the sake of simplicity. As an overview of the experimental methodology, it can be mentioned that the database has been divided into three parts: training, test, and validation sets such that all three sets have samples from each class). *PMI* and *MI* are used on training and test set to assess the importance of the features in regard to the class label. Some of these features selected by *PMI* and *MI* are qualitatively compared by visualizing their joint distributions with the class labels and as part of the comparisons, a permutation test is applied on the dataset to judge the robustness of both measures. Then, using the training set, an SVM is trained to predict the class label from the top features selected by *PMI* and *MI*. The parameters of SVM are optimized using its prediction accuracy on the test set. Then, the SVM is trained with these parameters on both training and test set together and its prediction accuracy is tested and reported on the validation set. Finally, the variables selected by the combined version of *PMI* with mRMR are used in the implementation of SVM and the results are presented.

4.1.1 Data Description

Arrhythmia dataset is used in this experimental study which is also available on the UCI machine learning archive (UCI 2005). Arrhythmias are disorders of the regular rhythmic beating of the heart. The aim of the dataset is to classify the sample in one of the 16 groups of arrhythmia of which class 1 means ‘normal’, classes 2 to 15 refer to different classes of arrhythmia, and class 16 refers to one of the unclassified arrhythmia types (Guvenir et al. 1997). The description of the dataset is tabulated in Table 4.1. The dataset contains 452 samples with 279 attributes, 206 of which are linear valued (the other 73 attributes/features are binary). The linear valued variables are discretized, but to avoid the problem mentioned in Section 3.1 and in Figure 3.3, 15 discrete levels were used. For discretization, for each feature, its mean μ and its standard deviation σ were used as in Peng, Long, and Ding (2005). The feature values between $\mu-\sigma/2$ and $\mu+\sigma/2$ are converted to 0. The 7 intervals of size σ to the right of $\mu+\sigma/2$ are converted to discrete levels from 1 to 7 and the 7 intervals of size σ to the left of $\mu-\sigma/2$ are mapped to discrete levels from -1 to -7. Very large positive or negative feature values are truncated and discretized to ±7 appropriately.

The database was divided into three parts. 50% of the data samples were used in each class for training set, 25% of the data samples in each class for the test set, and the remaining 25% of the data samples in each class for the validation set.

Table 4.1 : Arrhythmia Dataset Description

Class #	Class Name	# Total	# Training	# Test	# Validation
1	Normal	245	123	61	61
2	Ischemic Changes (Coronary Artery Disease)	44	22	11	11
3	Old Anterior Myocardial Infarction	15	8	4	3
4	Old Inferior Myocardial Infarction	15	8	4	3
5	Sinus Tachycardy	13	7	3	3
6	Sinus Bradycardy	25	13	6	6
7	Ventricular Premature Contraction (PVC)	3	1	1	1
8	Supraventricular Premature Contraction	2	1	1	0
9	Left Bundle Branch Block	9	5	2	2
10	Right Bundle Branch Block	50	25	13	12
11	1. Degree Atrioventricular Block	0	0	0	0
12	2. Degree AV Block	0	0	0	0
13	3. Degree AV Block	0	0	0	0
14	Left Ventricule Hypertrophy	4	2	1	1
15	Atrial Fibrillation or Flutter	5	3	1	1
16	Other	22	11	6	5

4.1.2 Feature Ranking/Selection

MI is successful in finding the important features which is predictive about the classes that have enough samples. However, in many scientific experiments, there are only a few samples of some classes. If a feature is important only in the prediction of a class with limited samples, *MI* gives a small score for that variable. Thus, that feature may be overlooked although it carries directly predictive information about a rarely seen class type.

Figure 4.1 shows the *PMI* versus *MI* scores of the features of arrhythmia dataset which shows that the measures present somewhat proportional values for many features. However, it must be also noted that the plot slants to the right. In other words, some features with the similar *MI* score can have various, a wider spectrum of, *PMI* scores (i.e. features with higher *PMI* is expected to have more mutually predictable relations with the class type). For an example of interesting findings, consider the features that have been circled in red (one of which is feature 125) and the feature marked with the red square (feature 267) in Figure 4.1. All these features have approximately the same *MI* scores (around 0.35) but some of them must be inferentially more important if they have higher *PMI* scores. As will be shown in the next subsections, these features could be more useful if included in future studies, in contrast to those with the same level of *MI* but lower *PMI* scores (e.g. feature 267). This is, of course, not to say that *MI* score is not important; however, in datasets with tens of thousands of variables, *PMI* would help as a valuable additional sort key along with *MI* because *PMI* weighs the mutual predictability, and thus, elaborates the *MI* measurement.

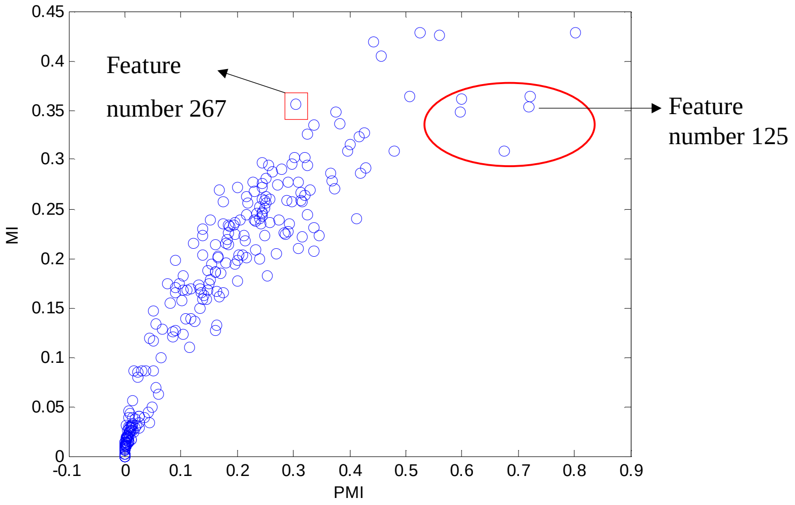

Figure 4.1 : *PMI* scores versus *MI* scores of the features of the arrhythmia dataset

4.1.3 Qualitative Comparison of Selected Features using Data Visualization

Figure 4.2 shows the plot of feature 125 versus the class label of the samples (just as in Figure 3.1, to be able to reflect the density of the points in 2D, small random noise is added to the feature's value rather than making it a 3D plot for the joint PDF; the same treatment is done to Figure 4.3 as well). It is shown that class label is 9 with 85% probability when this feature's value is higher than 2. This shows that features like 125 give information about rare but important events which *PMI* is more successful to identify than *MI* because of its formulation. Besides, knowing that feature 125 is important in the prediction of class 9 is valuable for the scientists who make researches in related fields.

Figure 4.3 shows the plot of feature 267. Feature 267 is one of the very top features in relevance using *MI* (it is also among top by *PMI*). The quaintness that the plot of feature 125 has cannot seem to be present in this one.

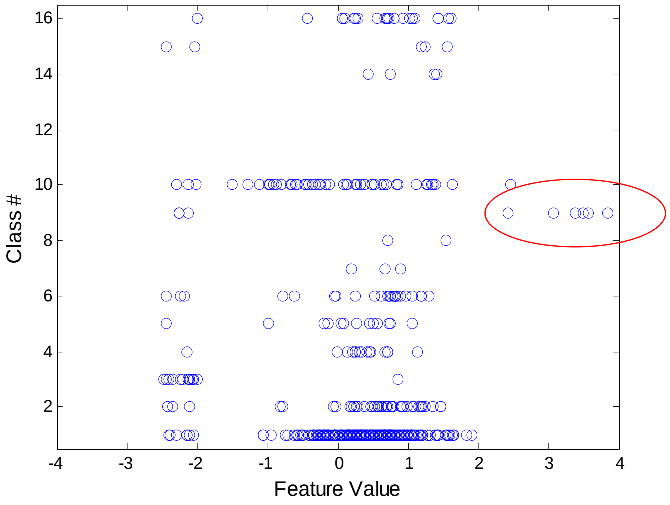

Figure 4.2 : Plot of feature 125 versus the class label

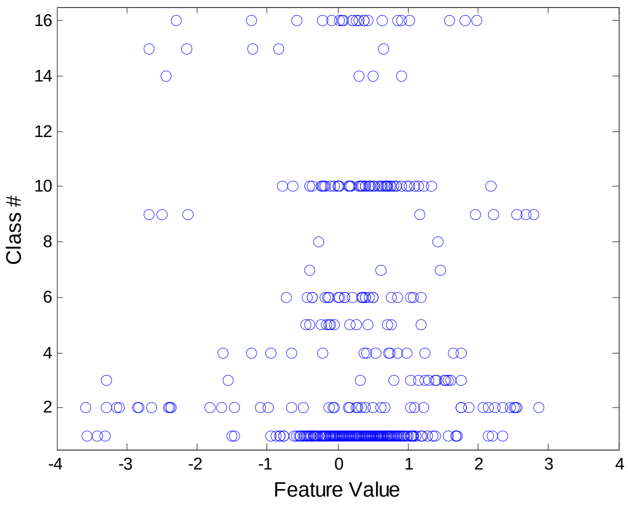

Figure 4.3 : Plot of feature 267 versus the class label

4.1.4 Quantitative Comparison of Selected Features using Permutation Test (Randomized Resampling)

In this section, a permutation test (Good 1994) is applied for testing the robustness of *MI* and our proposed measure *PMI*. For this purpose, all the features (not the class labels) of arrhythmia dataset have been randomly shuffled (i.e. each feature is randomly resampled using the values it takes in the dataset). If the dataset is considered as a matrix, with each column as a feature and each row as a data sample, then this process is basically randomly shuffling each column (independently). This process most likely destroys the relations between the features and the class labels. Then, *MI* and *PMI* scores of the features of the shuffled dataset are recomputed. It is expected that the features on the shuffled dataset have smaller scores than they have on the original. For each measure, the ratio of the sum of scores was calculated on the original dataset to the sum of the scores obtained on the shuffled one. This ratio, that we called original to noise ratio, is used to express the robustness of the two mutual information measures. Since the process involves randomness, it has been performed 100 times for statistical significance. The results of the 100 trials are shown in Figure 4.4. Based on these results, we can conclude that *PMI* is more successful in distinguishing between the real sampled dataset and noisy (shuffled) dataset because the original to noise ratio of *PMI* is higher than of *MI* in every trial.

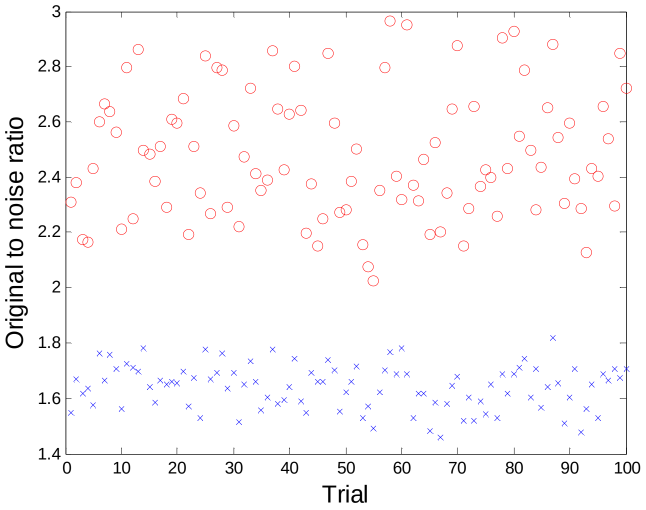

Figure 4.4 : Original to noise ratio. Red circles denote the original to noise ratio of *PMI* and blue x-marks denote the original to noise ratio of *MI*

4.1.5 Quantitative Comparison of Selected Features using Support Vector Machines

Features selected by *MI* and our proposed *PMI* measures need to be tested on how much joint predictive power they have of the target (class labels). To test this, a popular machine learning tool Support Vector Machine (SVM) is used described in section 2.1.2.

As shown in Table 4.1, arrhythmia dataset contains 452 samples, 245 of which are ‘normal’ class type. The dataset is divided into three groups: 50% for train, 25% for test and 25% for validation. The distribution of the samples to the datasets has been done so as that each set contains samples from class types with the above mentioned percentages. Table 4.1 shows the class names and number of samples of each dataset for each class.

As mentioned in Section 4.1.2, the features are ranked by proposed *PMI* (the experiments were repeated in the same manner for *MI*). Using the results reported by Peng, Long, and Ding (2005) and using sequential backward selection (Bishop 1995) empirically, it has been determined that in the order of 30 features are required for the best performance of SVM. To compensate the redundant features (which is very typical of this dataset), it has been concluded that 40 features would be a good approximation to the optimal number of top ranking features to use in the subsequent studies. Then, an SVM was trained on the training set using various settings for the SVM parameters. Obtained models were applied to the test set and best fitted SVM parameters have been determined (these are *C* is 3 and use *g*, or in some texts σ, as default). The SVM is trained one last time using the best settings and using both the training and the test sets. It is tested on the left-aside validation set. Although 1.0 is a good default setting, different values of α and β have been used in the experiments. The results are shown in Tables 4.2 and 4.3 for the test and the validation sets, respectively. Both 30 and 40 top features have been tried (*N* denotes this number in the tables).

Table 4.2 : SVM Results on the Test Set

	N=40, α=1, β=1		N=40, α=0.5, β=1		N=30, α=1, β=1	
	PMI	***MI***	***PMI***	***MI***	***PMI***	***MI***
Class #						
1 (61)	0.89	0.92	0.92	0.92	0.85	0.93
2 (11)	**0.73**	**0.64**	**0.82**	**0.64**	0.82	0.82
3 (4)	0.50	0.75	0.75	0.75	**0.75**	**0.50**
4 (4)	**0.75**	**0.25**	**0.50**	**0.25**	**0.50**	**0.25**
5 (3)	**0.33**	**0**	0	0	0.33	0.33
6 (6)	0	0	0	0	0	0
7 (1)	0	0	0	0	0	0
8 (1)	0	0	0	0	0	0
9 (2)	0.50	0.50	0.50	0.50	0.50	0.50
10 (13)	0.46	0.77	**0.85**	**0.77**	**0.46**	**0.38**
11 (0)	-	-	-	-	-	-
12 (0)	-	-	-	-	-	-
13 (0)	-	-	-	-	-	-
14 (1)	0	0	0	0	0	0
15 (1)	0	0	0	0	0	0
16 (6)	0	0	0	0	0	0
Overall	**0.66**	**0.68**	**0.72**	**0.68**	**0.65**	**0.67**

Table 4.3 : SVM Results on the Validation Set

	N=40, α=1, β=1		N=40, α=0.5, β=1		N=30, α=1, β=1	
	PMI	***MI***	***PMI***	***MI***	***PMI***	***MI***
Class #						
1 (61)	0.93	0.95	0.95	0.95	0.93	0.95
2 (11)	0.55	0.55	0.55	0.55	0.55	0.67
3 (3)	1.00	1.00	1.00	1.00	1.00	1.00
4 (3)	**1.00**	**0.33**	**0.67**	**0.33**	**1.00**	**0**
5 (3)	1.00	1.00	1.00	1.00	1.00	1.00
6 (6)	0.33	0.50	0.50	0.50	0.17	0.50
7 (1)	0	0	0	0	0	0
8 (0)	-	-	-	-	-	-
9 (2)	1.00	1.00	1.00	1.00	1.00	1.00
10 (12)	**0.50**	**0.42**	0.42	0.42	**0.42**	**0.33**
11 (0)	-	-	-	-	-	-
12 (0)	-	-	-	-	-	-
13 (0)	-	-	-	-	-	-
14 (1)	0	0	0	0	0	0
15 (1)	0	0	0	0	0	0
16 (5)	0	0	0	0	0	0
Overall	**0.76**	**0.75**	**0.76**	**0.75**	**0.74**	**0.74**

MI and proposed *PMI* give similar classification accuracies in the prediction of the classes which have large number of samples such as class #1 (the ‘normal’ class). However, according to the obtained results, it can be concluded that proposed *PMI* surpasses *MI* in selecting predictive features which have valuable information about rare but important events (shown in bold in the Tables 4.2 and 4.3). This advantage of *PMI* over *MI* could make it a useful data mining tool for the researchers in fields, such as in the biomedical field, in which datasets might contain many samples about one class but only a few about some other important classes.

Also note that the quality of the results obtained in this subsection is limited with the learning capacity of SVMs. Even though, the features that capture suspicious coincidences are presented to SVM, it may not (be able to) take such relations into account in the learning process due to the curse of dimensionality.

4.1.6 Quantitative Comparison of Selected Features Combining With mRMR

PMI is combined with mRMR in Section 3.2 with the aim of selecting a minimal subset that represents the problem. An SVM is trained again with the same methodology (also same SVM parameters determined are used) in Section 4.1.5 using the selected variables of combined version of *PMI* with mRMR. The results are shown in Table 4.4 for the validation set.

Table 4.4 : SVM results on the validation set using the combined version of *PMI* with mRMR

	N=20, α=1, β=1		N=30, α=1, β=1	
	PMI	***PMI* with mRMR**	***PMI***	***PMI* with mRMR**
Class #				
1 (61)	0.95	0.92	0.93	0.95
2 (11)	0.55	0.55	0.55	0.55
3 (3)	1.00	1.00	1.00	1.00
4 (3)	0.67	1.00	1.00	1.00
5 (3)	1.00	1.00	1.00	1.00
6 (6)	0.50	0.33	0.17	0.33
7 (1)	0	0	0	0
8 (0)	-	-	-	-
9 (2)	1.00	1.00	1.00	1.00
10 (12)	0.17	0.33	0.42	0.50
11 (0)	-	-	-	-
12 (0)	-	-	-	-
13 (0)	-	-	-	-
14 (1)	0	0	0	0
15 (1)	0	0	0	0
16 (5)	0	0	0	0
Overall	**0.73**	**0.73**	**0.74**	**0.77**

As seen in the results, best prediction accuracy (0.77) is obtained with 30 variables selected by the combined version of *PMI* with mRMR.

4.2 SO_2 DATASET

Sulfur dioxide (SO_2) is an issue of increasing public concern due to its recognized adverse effects on human health. Therefore, accurate SO2 prediction models are very important tools in developing public warning strategies. A comparison of *PMI* and *MI* is included in this experimental study which shows that *PMI* can help detecting important features (according to the specialists in Environmental engineering) that can be missed by *MI*.

4.2.1 Dataset Description

The air pollutant parameter measurements used in this study were procured from the Director of Istanbul Metropolitan Municipality Environment Protection and Control Office which has 10 automatic air quality measuring stations in Istanbul, Turkey, to observe the air pollution in the atmosphere of Istanbul continuously. These measurements have been observed at 15 min interval. Our dataset contains the measurements of two of these stations, Kadikoy and Sarachane, from July 2003 to June 2004. The reason for choosing these two locations and that time period is that they contain less missing values than the other stations. Therefore, meteorological variables were chosen from Florya and Goztepe meteorological stations of Government Meteorology Works Office which are the nearest stations to Kadikoy and Sarachane, respectively. Meteorological parameters are continuously saved in 17 stations of Government Meteorology Works Office at 1 hour interval. The data from the Asian Side and European Side contain 324 and 261 samples, respectively, after removing the samples with missing values. The included air pollutant parameters are daily average concentration of SO2, nitrogen oxide (NO), nitrogen dioxide (NO2), total hydrocarbons (THC), dust, ozone (O_3), and daily maximum SO2 concentration. Meteorological parameters are daily average outdoor temperature (OT), average cloudiness (C), average relative humidity (RH), average pressure (P), total amount of solar radiation (SR), average wind speed (WS), and total amount of rain (R). Target variable is next day's daily maximum SO_2 concentration. Table 4.5 shows these variables' statistical parameters and Figure 4.5 shows the plot of maximum SO_2 concentration at time t+1 versus each input variable's value at time t.

Table 4.5 : Statistical parameters of the SO_2dataset

Variable	Minimum	Maximum	Average	Standard
Average SO_2 ($\mu g/m^3$)	0	82	16.102	13.068
Maximum SO_2 ($\mu g/m^3$)	0	188	39.859	31.987
NO ($\mu g/m^3$)	3	587	46.551	63.169
NO_2 ($\mu g/m^3$)	13	158	53.898	24.866
O_3 ($\mu g/m^3$)	0	86	14.45	10.535
Dust ($\mu g/m^3$)	9	191	55.662	29.21
Hydrocarbon ($\mu g/m^3$)	162	4091	1588.755	419.33
Temperature (^oC)	-5.3	28.8	13.509	7.686
Wind Speed (m/s)	0.4	7.1	2.537	1.137
Solar Radiation (Hour)	0	13.2	6.086	4.284
Cloudiness (0 – 10)	0	10	4.855	3.431
Pressure (mbar)	988.2	1032.1	1012.452	6.514
Relative Humidity (%)	45.7	95.7	73.032	11.058
Rain (mm)	0	48.6	1.91	4.994

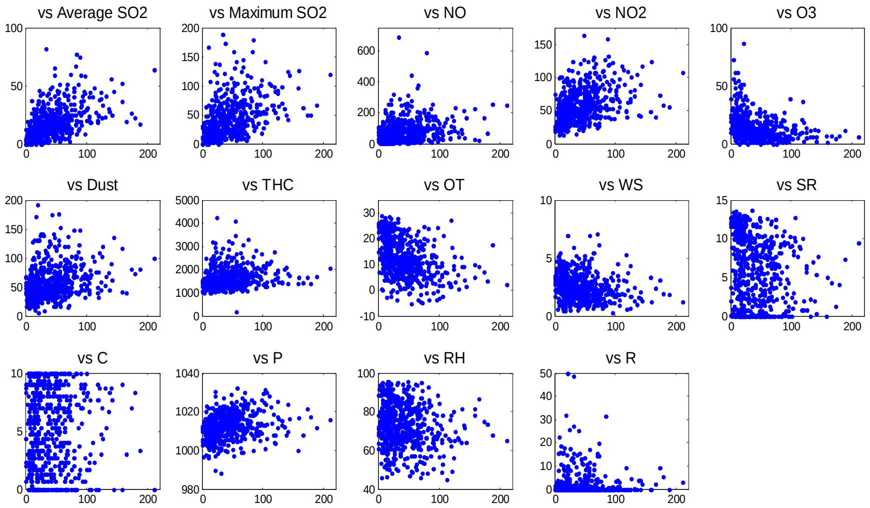

Figure 4.5 : Plot of maximum SO2 concentration at time t+1 versus each input variable's value at time t

All the variables of our dataset are linear valued, so we discretized them to calculate *MI*. For discretization, for each feature, we used its mean μ and its standard deviation σ.

The feature values between $\mu-\sigma/2$ and $\mu+\sigma/2$ are converted to 0. Since we used 9 discrete levels, the 4 intervals of size σ to the right of $\mu+\sigma/2$ are converted to discrete levels from 1 to 4 and the 4 intervals of size σ to the left of $\mu-\sigma/2$ are mapped to discrete levels from -1 to -4. Very large positive or negative feature values are truncated and discretized to ±4 respectively.

4.2.2 Feature Ranking/Selection

PMI and *MI* scores of all the features with the target variable (next day's maximum SO2 concentration) were calculated to measure their relevance. In order for *MI* and *PMI* scores to make intuitive sense, we normalized them by dividing them with the entropy of the target variable (as in Eq. 2.11). *PMI* and *MI* scores and their normalized values are shown in Table 4.6. Figure 4.6 shows the *PMI* versus *MI* scores.

Table 4.6 : *PMI* and *MI* scores of the input variables

Input variables	*PMI*	Normalized *PMI*	*MI*	Normalized *MI*
Average SO_2 concentration	0.5171	27.54%	0.2882	15.35%
Maximum SO_2 concentration	0.4509	24.01%	0.2303	12.27%
Average outdoor temperature	0.5272	28.07%	0.1830	9.74%
Average NO_2 concentration	0.4936	26.28%	0.1690	9.00%
Average O_3 concentration	0.3212	17.10%	0.1211	6.45%
Average wind speed	0.2661	14.17%	0.1108	5.90%
Average NO concentration	0.2416	12.86%	0.1020	5.43%
Average pressure	0.2559	13.63%	0.0932	4.96%
Dust	0.2467	13.14%	0.0739	3.93%
Total hydrocarbons	0.1546	8.23%	0.0563	3.00%
Total amount of solar radiation	0.2649	14.10%	0.0554	2.95%
Relative humidity	0.1064	5.66%	0.0451	2.40%
Total amount of rain	0.0141	0.75%	0.0299	1.59%
Average cloudiness	0.0432	%2.30	0.0152	0.81%

As seen in Figure 4.5, the measures present somewhat proportional values for many features. However, as it was in arrhythmia dataset, the plot slants to the right. In other words, some features with the similar *MI* score can have various, a wider spectrum of, *PMI* scores (i.e. features with higher *PMI* is expected to have more mutually predictable relations with the class type). The most interesting finding is the solar radiation which is marked with the red square. It is in 11th order according to *MI* measure with a normalized value of 2.95 which seems irrelevant with the target variable. However, it is ranked 7th by *PMI* measure with a normalized value of 14.10 which means that 14.10% of target variable's uncertainty can be removed by knowing the actual values of solar radiation.

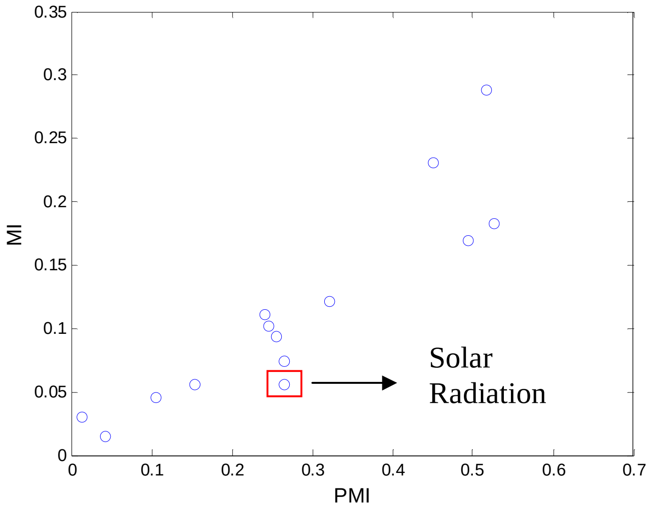

Figure 4.6 : *PMI* scores versus *MI* scores of the features of the SO_2 dataset

It must be noted that this finding coincidences with the fact (according to the specialists and existing studies in environmental engineering) that outdoor temperature is the most important variable among meteorological variables (a key factor) in the prediction of SO_2 concentration and solar radiation is directly related with outdoor temperature.

5. CONCLUSIONS

Feature selection is a very important computational preprocessing step for most subsequent research, such as building accurate inferential or classification models of the observed phenomenon, or elaborating the found dependencies (i.e. connections of the selected features to the target functions), perhaps even by means of laboratory/experimental studies. However, feature selection is not an easy task, especially when the data is of high dimensionality and the relations are multivariate and nonlinear, and when there are many factors that weakly contribute to the target functions.

Shannon's Mutual Information (*MI*) is a well known information theoretic measure of dependence which has recently been used for feature selection as a filter. However, mutual information works effectively for the well sampled datasets. Thus, in many areas of experimental sciences, it is a difficult task to calculate mutual information accurately due to the limited sample size. Moreover, mutual information measure is not very sensitive to rare but predictive relations as is, because such rare relations may have little overall information content from information theoretic perspective.

In this thesis, firstly, it is showed that using Shannon's mutual information measure can lead to missing relations that can be very predictive of rare but well-predictable classes. Then, the mutual information measure is developed into a novel measure of dependence, Predictive Mutual Information (*PMI*), by the use of the concept of suspicious coincidences (predictable relations). The proposed measure weighs more the samples with predictive powers, thus effectively eliminates the samples with no predictive contribution. This modification makes *PMI* take the suspicious coincidences also into account. Thus, *PMI* works better for databases involving possibly rare but well-predictable classes and also overcomes the low sampling rate problem in a rather indirect way. With the aim of selecting a more compact and discriminative subset of variables, *PMI* is combined with mRMR (Peng, Long, and Ding 2005) approach which avoids selecting redundant variables.

The usefulness of *PMI* and superiority over *MI* are demonstrated on both toy and real datasets. In conclusion, we believe that *PMI* measure could be a more useful measure than Shannon’s mutual information measure under the conditions typical of real-world datasets (such as in biomedical), in which limited number of observations are available especially from some important classes (phenotypes) of the target function to predict. Because, unlike *MI*, *PMI* is not just about transferring bits over a noisy channel. It has a goal of detecting orderly relations. Thus, it also helps keep the rare but well-predictable classes in the calculations without having their effect blurred in otherwise random relations. This is, of course, not to conclude that *MI* is not important; however, in datasets with tens of thousands of variables, *PMI* would help as a valuable additional sort key along with *MI*. That is, since *PMI* weighs the mutual predictability, it elaborates the *MI* measurement and helps avoid false negatives.

REFERENCES

Books

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* Oxford University Press, Oxford.

Good P., 1994. *Permutation Tests,* Springer, New York.

MacKay, D., 2003. *Information theory, Inference, and Learning Algorithms,* Cambridge University Press.

Scholkopf B., Smola A., 2002. *Learning with Kernels*, MIT Press, Cambridge, MA.

Shawe-Taylor J., Cristianini N., 2004. *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Vapnik, V., 1995. *The Nature of Statistical Learning Theory*, Springer, New York.

Vapnik, V., 1998. *Statistical Learning Theory*. New York: Wiley.

Periodical Publications

Baofeng G. & Nixon M.S., 2007. Gait Feature Subset Selection by Mutual Information, *Biometrics: Theory, Applications, and System*, BTAS , pp. 1-6.

Becker, S., 1996. Mutual Information Maximization: Models of Cortical Self-Organization, *Network*, 7(1) pp. 7-31.

Breiman, L. & Friedman, J. H., 1985. Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, 80, pp. 580-598.

Burges, C.J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition, *Knowledge Discovery and Data Mining*, 2, pp. 1-43.

Ding, C. & Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology*, 3(2), pp. 185-205.

Endres, D. & Földiák P., 2005. Bayesian Bin Distribution Inference and Mutual Information, *IEEE Transactions on Information Theory*, 51(11), pp. 3766-3779.

Favorov, O.V. & Ryder, D., 2004. SINBAD: a neocortical mechanism for discovering environmental variables and regularities hidden in sensory input, *Biological Cybernetics*, 90, pp. 191-202.

Guyon, I. & Elisseeff, A., 2003. An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3, pp. 1157-1182.

Hsu, C.W. & Lin, C.J., 2002. A Comparison of Methods for Multi-Class Support Vector Machines, *IEEE Trans. Neural Networks*, 13, pp. 415-425.

Ioannidis J.P.A., 2005. Why most published research findings are false? *PLoS Med*, 2(8), pp.124.

Jiang, D., Zhang, Y., Hu, *X*., Zeng, Y., Tan, Jianguo, Shao, D., 2004, Progress in developing an ANN model for air pollution index forecast, *Atmospheric Environment*, 38, pp. 7055-7064.

Joe, H., 1989. Relative Entropy Measures of Multivariate Dependence, *Journal of the American Statistical Association*, 84(405), pp. 157-164.

Kurşun, O. & Favorov, O., 2004. SINBAD Automation of Scientific Discovery: From Factor Analysis to Theory Synthesis. *Natural Computing*, 3(2), pp. 207-233.

Kwak, N. & Choi, C.H., 2002. Input Feature Selection by Mutual Information Based on Parzen Window, *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(12), pp. 1667-1671.

Mjolsness, E. & DeCoste, D., 2001. Machine learning for science: state of the art and future prospects, *Science*, 293, pp. 2051-2055.

Novoviccová, J., Somol, P., Haindl, M., Pudil, P., 2008. Conditional Mutual Information Based Feature Selection for Classification Task, *Progress in Pattern Recognition, Image Analysis and Applications*, 4756, pp. 417-426.

Peng, H., Long, F., Ding, C., 2005. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), pp. 1226-1238.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods, *Journal of Machine Learning Research*, 3, pp 371-1382.

Shannon, C. E., 1948. A mathematical theory of communication, *Bell System Technical Journal*, 27, pp. 379-423, pp. 623-656.

Spellman, G., 1999. An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom. *Applied Geography*, 19, pp. 123-136.

Stigler, S.M., 1968. Francis Galton’s Account of the Invention of Correlation, *Statistical Science*, 4(2), pp. 73-86.

Wang Z, Massimi C D, Tham M T , Morris A J, 1994. A procedure for determining the topology of multilayer feed-forward neural networks, *Neural Networks,* 7, pp. 291-300.

Warner, B., & Misra, M., 1996. Understanding neural networks as statistical tools. *The American Statistician*, 50, pp. 284-294.

Zhang,J.G. & Deng, H.W., 2007. Gene selection for classification of microarray data based on the Bayes error, *BMC Bioinformatics*, 8, pp. 370.

Other Publications

Guvenir, H.A., Acar, B., Demiroz G., Cekin, A., 1997. A Supervised Machine Learning Algorithm for Arrhythmia Analysis, *Proceedings of the Computers in Cardiology Conference*, Lund, Sweden.

Caruana, R. Lawrence, S. Giles, L., 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping, *Neural Information Processing Systems*, Denver, Colorado.

Rumelhart DE, Hinton GE, Williams RJ, 1986. Learning internal representations by error propagation. In: Rumelhart DE, Mcclelland JL, *PDP Research Group (eds) Parallel Distributed Processing: Explorations in the Microstructure Of Cognition*, MIT Press, Cambridge, Mass, 1, pp. 318-362.

UCI Learning Repository, 2005, http://www.ics.uci.edu/mlearn/ MLSummary.html.

Chih-Chung Chang & Chih-Jen Lin, 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

VITAE

Name Surname	: Cemal Okan ŞAKAR
Address	: Bahçeşehir Üniversitesi Mühendislik Fakültesi Çırağan Cd. Osmanpaşa Mektebi Sk. No: 4 – 6 34349 Beşiktaş / İstanbul / Türkiye
Birth Place / Year	: Erzurum - 1984
Languages	: Turkish (native) - English
Elementary School	: İstiklal Primary School – 1995
High School	: Hüseyin Yıldız Anatolian High School - 2002
BSc	: Yıldız Technical University - 2006
MSc	: Bahçeşehir University – 2008
Name of Institute	: Institute of Science
Name of Program	: Computer Engineering
Publications	: **Şakar C.O**., & Kurşun, O., Predictive Mutual Information and Its Use In Feature Selection, *Computational Statistics & Data Analysis*, submitted (2008). **Şakar C.O**., Albayrak, Ş., Demir, G., Ozdemir, H., Yalçın, Ş., Determining the importance of input variables in the prediction of tropospheric ozone concentration. *Advances in atmospheric sciences*, submitted (2007).
Work Experience	: Bahçeşehir University Software Engineering Department Teaching and Research Assistant (August 2006 – Today)