A Mutual Information estimator for continuous and discrete variables applied to Feature Selection and Classification problems

Frederico Coelho [1] , **Antonio P. Braga** [2] , **Michel Verleysen** [3]

[1] *Universidade Federal de Minas Gerais*
Brazil
E-mail: fredgfc@ufmg.br

[2] *Universidade Federal de Minas Gerais*
Brazil
E-mail: apbraga@ufmg.br

[3] *Université Catholique de Louvain*
Belgium
E-mail: michel.verleysen@uclouvain.be

Received 23 March 2015

Accepted 11 April 2016

Abstract

Currently Mutual Information has been widely used in pattern recognition and feature selection problems. It may be used as a measure of redundancy between features as well as a measure of dependency evaluating the relevance of each feature. Since marginal densities of real datasets are not usually known in advance, mutual information should be evaluated by estimation. There are mutual information estimators in the literature that were specifically designed for continuous or for discrete variables, however, most real problems are composed by a mixture of both. There is, of course, some implicit loss of information when using one of them to deal with mixed continuous and discrete variables. This paper presents a new estimator that is able to deal with mixed set of variables. It is shown in experiments with synthetic and real datasets that the method yields reliable results in such circumstance.

Keywords: Feature Selection; Mutual Information; Classification.

1. Introduction

Mutual information (MI) [1] has been applied to a wide range of machine learning problems [2,3]. It is a well established approach, especially for estimating uni and multi-variate non-linear relations, being also applied in the context of Feature Selection (FS) [4]. In order to evaluate mutual information, densities of dependent and independent variables should be estimated. In practice, evaluating mutual information is not straightforward, since it requires *a priori* knowledge of densities, however, not much information about generator functions is available in advance, what requires an estimator to be adopted.

A MI estimator for classification problems is derived from the Kraskov estimator [5], then developed by Goméz et al. [6]. It addresses classification tasks by using the discrete nature of the output variable, and can also be applied to multi-class feature selection problems. Nevertheless, like the original Kraskov estimator, this approach is also restricted to continuous input variables. However, most real-domain applications contain not only continuous but also discrete variables, which are usually treated separately. In mixed variable problems, usually continuous features are discretized and then density estimators for discrete variables are used in order to evaluate MI.

Other strategies can be found in the literature to estimate mutual information, like in [7] where Parzen-window is used, but without specifying whether there was differential treatment for discrete and continuous variables. In [8] a filter method uses the Fraser estimator extended version to estimate densities of continuous variables and contingency tables to discrete ones. This paper aims to propose a new estimator based on the original Kraskov method that is able to deal concurrently with continuous and discrete variables in order to perform feature selection. The ability to aggregate discrete and differential entropies is the main aspect of the proposed estimator. This is based on the fact that differential entropy of a random variable and entropy of its discretized version are different. Loss of information related to relevant features may appear when discretization methods are applied. The proposed method yields improved performance in such cases, since discrete and continuous variables are jointly considered.

Experiments with datasets composed by mixed set of variables are carried out for feature selection problems.

2. Mixed Entropy and Mutual Information

Common approaches for MI estimation include dynamic allocation of histogram bins [4], recursive partitioning of the input domain [9] and kernel density estimators [10]. Nevertheless, as detailed below, MI estimation from discretized variables are shifted from the estimation obtained directly from the original continuous variables.

Consider a continuous random variable Z with a continuous probability density function $f(z)$ and that the space of Z is discretized into fixed intervals Δ and each interval i is defined as $[i\Delta,(i+1)\Delta]$. For each interval i, as a direct consequence of the *mean value theorem*, it is possible to find a value z_i for which

$$f(z_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(z)dz\ . \qquad (1)$$

A discrete random variable Z^Δ can be defined over a countable number of values z_i, being one per interval i of Z. In this case the probability p_i associated to z_i can be written on the basis of the probability density function of Z as $p_i = f(z_i)\Delta$. Cover and Thomas [1] show that the discrete entropy of the quantized variable Z^Δ is given by

$$\begin{aligned} H(Z^\Delta) &= -\sum_{-\infty}^{\infty} p_i\ log\ p_i & (2)\\ &= -\sum \Delta f(z_i)\ log f(z_i) - log\Delta & (3) \end{aligned}$$

if $\sum f(z_i)\Delta = \int f(z) = 1$.

It can be shown [1] that the first term in Equation 3 tends to the integral of $-f(z)\log f(z)$ as $\Delta \to 0$, if $f(z) log\ f(z)$ is Riemann integrable. This implies that the entropy of the discrete random variable Z^Δ and the differential entropy of the continuous random variable Z relate as

$$H\left(Z^\Delta\right) + log\Delta \to h(Z)\ as\ \Delta \to 0\ ; \qquad (4)$$

see theorem 8.3.1 in [1].

Equation 4 shows that the entropies of the original continuous variables and their discretized versions are not the same, what suggests that a specific estimator for mixed variables is needed.

For instance (see [1]), if $Z \sim \mathcal{N}\left(0,\sigma^2\right)$ with $\sigma^2 = 100$, will be necessary, on the average, $n + \frac{1}{2}\log\left(2\pi e\sigma^2\right) = n + 5.37$ bits to describe Z to n bit accuracy.

2.1. *Entropy of a mixed set of variables*

Given a discrete random variable X and a continuous variable Z, the mixed joint entropy $\mathcal{H}(Z,X)$ can be formulated as

$$\mathcal{H}(Z,X) = H(X) + h(Z\,|\,X)\ , \qquad (5)$$

where $H(X)$ is the entropy of a discrete random variable and $h(Z \mid X)$ is the differential conditional entropy of a continuous variable Z.

It is important to notice that in Equation 5 the entropy is the sum of two different quantities: the differential entropy of a continuous variable and the discrete entropy of a discrete one. Since the random variable X is discrete the conditional differential entropy in Equation 5 is given by

$$h(Z \mid X) = \sum_{x \in X} p(X = x)\, h(Z \mid X = x) \ . \tag{6}$$

Then, the mixed entropy of a discrete random variable X and a continuous one Z can be formulated as

$$\begin{aligned}\mathscr{H}(Z,X) &= H(X) + \sum_{x \in X} p(X = x)\, h(Z \mid X = x) \\ &= H(X) - \sum_{x \in X} p(X = x) \\ &\quad \int_S f(Z \mid X = x)\, log f(Z \mid X = x)\, dz\end{aligned} \tag{7}$$

where S is the support set of the random variable Z.

2.2. *Mutual Information between a mixed set of variables and a discrete one*

Let us now consider V as a random variable set composed by a discrete random variable X and a continuous random variable Z, such as $V = \{X \cup Z\}$ and also considering to another discrete random variable Y

The MI between V and Y can be defined, in terms of the mixed entropy, as $MI(V,Y) = \mathscr{H}(V) - \mathscr{H}(V \mid Y)$ that can be rewritten as

$$\begin{aligned}MI(V,Y) = H(X) + \sum_{x \in X} p(X = x)\, h(Z \mid X = x) - \sum_{y \in Y} p(Y = y) \\ \left(H(X \mid Y = y) + \sum_{x \in X} p(X = x \mid Y = y)\, h(Z \mid X = x, Y = y) \right)\end{aligned} \tag{8}$$

3. Mixed Mutual Information Estimator

An estimator of the Mixed Mutual Information (MMI) can be developed by replacing the differential entropy quantities in Equation 8 by the Kozachenko-Leonenko entropy estimator

$$\widehat{h}(Z) = -\psi(k) + \psi(N) + log\, C_d + \frac{d}{N} \sum_{n=1}^{N} log\, \varepsilon(n,k) \tag{9}$$

as presented in [5], where k is the number of nearest neighbors that should be set by the user, N is the number of patterns, d is the dimension of Z, C_d is the volume of the d-dimensional unitary sphere, $\psi(\cdot)$ is the digamma function and $\varepsilon(n,k)$ is twice the distance from z_n to its k^{th} neighbor.

Then, after some algebraic manipulations and generalizing for the case where $V = \{X,Z\}$, with $X = \{X_1, \ldots, X_n\}$ being a set of n discrete random variables, $Z = \{Z_1, \ldots, Z_t\}$ being a set of t continuous random variables and Y a discrete random variable, the Mixed Mutual Information estimator can be written as:

$$\begin{aligned}MI(V,Y) = {} & H(X_1) + \sum_{g=2}^{n} p\left(X_1, \ldots, X_{g-1}\right) H\left(X_g \mid X_{g-1}, \ldots, X_1\right) \\ & - \sum_{y \in Y} p(Y = y)\,[H(X_1 \mid Y = y)] - \sum_{y \in Y} p(Y = y) \\ & \left[\sum_{g=2}^{n} p\left(X_1, \ldots, X_{g-1} \mid Y = y\right) H\left(X_g \mid X_{g-1}, \ldots, X_1, Y = y\right) \right] \\ & + \sum_{x \in X} p(X = x)\, h(Z \mid X = x) - \sum_{y \in Y} p(Y = y) \\ & \sum_{x \in X \mid Y = y} p(X = x \mid Y = y)\, h(Z \mid X = x, Y = y) \ .\end{aligned} \tag{10}$$

The equation 10 for a set of continuous and discrete variables depends on the definition of the mixed entropy $\mathscr{H}$. This mixed entropy definition allows the use of different quantities as discrete entropy and differential entropy in the same framework. This is the key point of this new MI Estimator: the ability to sum, in a proper way, discrete and differential entropies.

4. Experiments

The experiments in this work show that there is some loss of information when discretizing continuous features in feature selection problems. This will be particularly noticed for those datasets whose most relevant features are continuous. Also the experiments will show that the MMI defined by Equation 10 is effective and consistent when applied to an information theoretic based feature selection procedure. The results obtained by the application of the MMI estimator are compared with a discrete approach that works by discretizing continuous variables. Firstly, feature selection using the MMI estimator was applied to all continuous and discrete features, generating the S_{mix} feature subset. In the second experiment, which will serve as a reference for comparison, continuous variables are discretized and an estimator of MI for discrete variables (based on histograms) is used.

A forward-backward sequential supervised feature selection algorithm [6,11] is implemented in the experiments as follows: during the forward step the selected feature subset S starts empty; at each iteration the feature f_i that together with S has the largest MI with Y, is permanently added to S. The procedure continues until a given stopping criterion is reached. The backward step starts from the final subset S of the forward step. At each iteration each selected feature f_j is individually and temporarily excluded from S (giving S_j) and the MI between S_j and Y is evaluated. The set S_j with the largest MI value is selected and, if S_j is more relevant than S given a stopping criterion (detailed below), then f_j is definitively excluded from S, otherwise the procedure is stopped.

Other forward-backward schemes could be adopted as well, however, as the goal of this paper is to evaluate the new MMI estimator, the experiments are restricted to a single choice of the forward-backward feature selection, as detailed above.

Each experiment (feature selection) is performed 10 times in a cross validation framework. The mean classification accuracies of each final selected feature subset are compared. Linear Discriminant Analysis Method (LDA)[12] is used to evaluate the classification accuracy due to its simplicity and robustness.

4.1. Statistical test

At the end, the Wilcoxon test is applied to evaluate if the accuracy of a classifier, trained with the set of features selected using different settings of MI estimators, are equivalent or not.

4.2. Stopping criterion

In the forward phase, considering f_i as a feature from the initial set F and S the selected feature subset, with $f_i \notin S$, and since S has one dimension less than $S \cup f_i$, the $MI(S \cup f_i, Y)$

value can not be compared directly to $MI(S,Y)$. Therefore a permutation test [13,14] is applied as a stopping criterion: from the set $S \cup f_i$, the feature f_i has its elements randomly permuted forming another set $S \cup f_i^p$, where f_i^p is the permuted version of feature f_i. This permutation generates a random variable with the same distribution of f_i, but that does not have any relation with the output Y (the corresponding values of Y are not permuted). Actually, adding a random variable to set S, in theory, does not improve nor degrades MI estimation between S and Y, but it increases the dimension of S in order to make it comparable to $S \cup f_i$. Therefore, if $MI(S \cup f_i, Y) > MI(S \cup f_i^p, Y)$ then $S \cup f_i$ is more relevant than S then f_i can be added to S and the process continues. Otherwise the forward process is halted and no more features are added to S.

The same principle is applied in the backward phase, however in a slightly different way. As before, it is not possible to compare the result of $MI(S,Y)$ with the result of $MI(S \setminus f_i, Y)$ in order to verify if there is an increase in relevance when feature f_i is removed from S, because the sets have different dimensions. Permuting the feature $f_i \in S$ transforms this feature in a random variable with no relation with Y, thus, a set without the influence of f_i but with the same dimension of S is generated. Now it is possible to assess if $S \setminus f_i$ is more relevant than S. If $MI(S,Y) < MI((S \setminus f_i) \cup f_i^p)$, f_i can be definitively removed from S and the process continues, otherwise the backward process is halted.

4.3. Datasets

In order to assess the performance of the proposed method in contrast with a discretization approach, 13 different datasets with different characteristics regarding size and number of discrete and continuous variables were selected. The datasets and their main characteristics are described next

- **DCbench** (DCB) is a synthetic dataset that was designed for testing the MMI estimator, since the relation between input features and the output variable is known and controlled. This dataset is composed by four discrete and six continuous features, that are sampled from different distributions. The DCbench dataset has 10.000 samples. The output results from a combination of three continuous features (X_1, X_2 and X_3) and two discrete ones (X_7 and X_8), in the following way:

$$Y = sign\left(tanh\left(X_1\right) + sin\left(X_2\right) + X_7 + X_8 + X_3\right) \ . \tag{11}$$

- **Boston Housing** (BOH) dataset [15] is composed of 506 samples with 13 features (3 discrete and 10 continuous). Originally the output variable of this dataset is the house prices, which is a continuous variable, however, here it is transformed into a classification problem by splitting the output into two classes: prices larger or smaller than a given threshold as in [16].
- **Page Blocks Classification** (PGBL) dataset [15], which is composed by 5473 samples with 10 features, being 6 discrete and 4 continuous.
- **Spambase** dataset (SPAM) [15], a dataset of e-mail spams with 4601 samples composed by 55 continuous and 2 discrete features.
- **Multi-feature digit** dataset (MFEAT) [15] consists of features of handwritten numerals (“0” to “9”) extracted from a collection of Dutch utility maps. It has 2000 samples (200 per class) with 190 continuous and 459 discrete features.
- **KDD Cup 1999 Data** [15] (KDD) from the *Third International Knowledge Discovery and Data Mining Tools Competition*. This dataset has originally 22 classes, but in order to accomplish the binary classification in this work, 600 samples from classes 10 (portsweep) and 11 (ipsweep) were selected for the tests. The dataset has 15 continuous and 26 discrete features.
- **Buzz in social media** dataset [15] (BUZZ), composed by 1000 samples with 77 features, of which 43 are discrete and 35 are continuous.
- **South African Heart** dataset [17] (SAH), composed by 462 samples with 9 features, being 4 discrete and 5 continuous.
- **QSAR biodegradation** dataset [15] (BIO), containing values for 41 features (molecular descriptors) used to classify 1055 chemicals, being 24 discrete and 17 continuous features.
- **Blog Feedback** dataset [15] (BLOG), composed by 1000 samples with 280 features, being 260 discrete and 20 continuous.
- **Australian Credit Approval** dataset [17] (ACA), composed by 690 samples with 14 features, being 11 discrete and 3 continuous.
- **Thyroid Disease** dataset [17] (THD),composed by 7200 samples with 21 features, being 15 discrete and 6 continuous.
- **Body Fat** dataset [18] (BFAT),composed by 252 samples with 14 features, being 1 discrete and 13 continuous.

5. Results discussion

Results are summarized in Table 1. S_{mix} is the feature subset selected using the MMI estimator considering all discrete and continuous features in the initial set, and S_{dd} is the set of selected features obtained when using a discrete MI estimator and considering all features as discrete (continuous features are discretized). $\overline{Acc}$ is the mean accuracy for 10 fold cross validation and σ is the yielded standard deviation.

Table 1. Mean accuracy of a LDA classifier after feature selection.

Problem	LDA accuracy ($\overline{Acc} \pm \sigma$)	
	S_{mix}	S_{dd}
DCB	0.9267±0.0082	0.8584±0.0111
PGBL	0.9011±0.0281	0.7955±0.0139
BUZZ	0.8300 ± 0.0465	0.6140 ± 0.0924
BFAT	0.8014 ± 0.0758	0.7500 ± 0.1095
SAH	0.6533 ± 0.0745	0.5928 ± 0.0573
BIO	0.7526 ± 0.0226	0.7118 ± 0.0407
KDD	0.9933±0.0111	0.9617±0.0249
MFEAT	0.9661±0.0246	0.9445±0.0102
BLOG	0.7080 ± 0.0489	0.6884 ± 0.0495
ACA	0.8551 ± 0.0468	0.8551 ± 0.0463
THD	0.9888 ± 0.0179	0.9999 ± 0.0020
BOH	0.8440±0.0518	0.8459±0.0635
SPAM	0.6740±0.0219	0.6727 ± 0.0206

It can be observed from the results of Table 1 that:

- When considering the MI estimator for a mixed set of variables resulted in a higher gain of classifier performance for DCB, PGBL, BUZZ, BFAT, SAH and BIO;
- The DCB database has relevant continuous variables that were selected when using the S_{mix} estimator. Variables 1 and 2 are continuous and are one of the five most relevant ones according to F-score and Relief [11]. This features were not selected when using the discrete estimator. Similarly, the database PGBL has continuous relevant variables that were selected when using MMI estimator.
- For the BFAT database the selection method using the MMI estimator was only slightly better, probably due to the selection of variables 6 and 7 that are continuous. Relief and F-score also indicated that these features are among the top 5 most relevant ones, but the method using the discrete estimator did not select them. The results obtained when applying the MMI estimator to SAH dataset was just slightly better, since it selected features 9 and 5; while when discretizing continuous variables, variable 9 was selected jointly with other ones not including variable 5. Variable 9, which is discrete, is the most relevant one according to Relief and F-score.
- For the BIO dataset Relief and F-score ranked the continous variable 39 as one of the top 3 most relevant features; the top 3 features selected by F-score are all continuous. In this case the MMI estimator provided some aditional information in order to slightly improve LDA performance in relation to the discrete estimator.
- For BLOG dataset F-score and Relief disagree about the most important variable, but for both the outcome is a descrete one. The variables selected when discretizing or not the continuous variables were all different but discrete ones, so the MMI estimator did not present any gain of performance for them. However, selected features were all different on each case what indicates that feature relevance and coupling are affected by discretization.
- In the case of the BUZZ dataset, the most relevant feature is discrete according to both relevance indexes. This feature was not selected when continuous features where discretized, probably because the relation between discrete and continuous features are affected by discretization, explaining MMI improved performance.
- The average performance of the final set of selected features using the proposed estimator is, in half of the cases, significantly better than the one achieved by the subset S_{dd}, which discretizes all continuous variables and adopts histogram-based estimation.

The Wilcoxon test was used to verify if accuracy results obtained using each set of selected features are similar or not. The null hypothesis was that different sets generate similar results. Comparing the results obtained using S_{mix} and S_{dd} the Wilcoxon test calculates a p-value of 0.0013. As the p-value was less than 5% then the null hypothesis is rejected indicating that the results are not similar and that S_{mix} has a better performance than S_{dd}.

Although there are studies in the literature using mutual information to select variables in some datasets used here [8] [7] [19], none compares specifically the use of estimators for continuous and discrete variables. Besides that, the main purpose of those works was to achieve a better classification accuracy, while our goal here is to show that the use of an estimator capable of handling a set of continuous and discrete variables can improve feature selection results. For this we needed to use a robust classifier that did not depend on initial conditions, so that the effect of jointly dealing with discrete and continuous variable could be observed.

6. Conclusions

The development of pattern classification and feature selection methods based on MI requires that a probability density function is estimated. Performance of the resulting model depends on the estimated function. The existence of continuous and discrete variables in most real problems imposes, however, an additional problem for density and MI estimation. The entropy of a continuous variable is related to the entropy of its discretized version as described in Equation 4. By itself, the discrepancy between the two values suggest that discretization should be avoided when evaluating MI. Therefore, when dealing with pattern classification or feature selection tasks based on MI, estimators based on discrete variables should be applied to datasets composed only by discrete variables and, estimators based on continuous variables should be applied to datasets composed only by continuous variables. Therefore, the use of a specific estimator designed to deal with datasets composed by discrete and continuous features is important and up to now, it was not addressed in a direct way in the literature. Of course one can split datasets into discrete and continuous sets, and use the proper estimator on each partition. However, coupling among discrete and continuous features will be disregarded. In this paper a MI estimator for a mixed set of variables was presented and applied to real datasets in a feature selection framework. The method was formally described and compared with other estimation approaches applied to feature selection classification problems.

The key point of this work is Equation 5 that defines the mixed entropy as a sum of two different quantities (discrete and differential entropies), allowing the development of the proposed mixed mutual information estimator. According to Equation 4 this approach avoids discretization inaccuracies what may result in improved performances of the feature selection methods, as confirmed by the experiments presented in this paper and by the Wilcoxon test results.

Our next step is to apply our new estimator to other feature selection methods. Also, as a continuation of our work we are interested in applying this estimator to Multiple Instance Learning problems, where distance and similarity measures are used to classify bags of samples. MI would be used to determine whether a sample belongs to a positive or negative bag. Furthermore, we have interest to apply the new estimator to other pattern recognition problems.

Acknowledgments

This work was developed with financial support from CAPES (process number 1456-10-5) and CNPq (process number 141818/2010-7).

References

1. T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, (1991).
2. C. Krier, D. François, F. Rossi, and M. Verleysen. Feature clustering and mutual information for the selection of variables in spectral data. *Neural Networks*, pages 25–27, (2007).
3. I. Quinzan, J. M. Sotoca, and F. Pla. Clustering-based feature selection in semi-supervised problems. *Intelligent Systems Design and Applications, International Conference on*, **0**:535–540, (2009).
4. R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, **5**(4):537 –550, jul (1994).
5. A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **69**(6 Pt 2), June (2004).

6. V. Gómez-Verdejo, M. Verleysen, and J. Fleury. Information-theoretic feature selection for functional data classification. *Neurocomputing*, **72**:3580–3589, October (2009).
7. G. Doquire and M. Verleysen. Feature selection with mutual information for uncertain data. In *Proceedings of the 13th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK'11, pages 330–341, Berlin, Heidelberg, (2011). Springer-Verlag.
8. P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, **20**(2):189–201, Feb (2009).
9. G.A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. on Information Theory*, **45**(4):1315 –1321, may (1999).
10. R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, **18**(suppl 2):S231–S240, October (2002).
11. I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, (2006).
12. T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag, (2001).
13. D. François, V. Wertz, and M. Verleysen. The permutation test for feature selection by mutual information. In *in: ESANN 2006, European Symposium on Artificial Neural Networks*, pages 239–244, (2006).
14. P Good. Permutation tests. *Technometrics*, **43**(June):114–114, (2008).
15. K. Bache and M. Lichman. UCI machine learning repository, (2013). Available at `http://archive.ics.uci.edu/ml`.
16. F. van der Heijden, R. Duin, D. de Ridder, and D. M. J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Wiley, 1 edition, nov (2004).
17. Knowledge extraction based on evolutionary learning repository. Available at `http://sci2s.ugr.es/keel/category.php?cat=clas`.
18. Statlib data. Available at `http://lib.stat.cmu.edu/datasets/`.
19. E. Schaffernicht and H. Gross. *Artificial Neural Networks and Machine Learning – ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part II*, chapter Weighted Mutual Information for Feature Selection, pages 181–188. Springer Berlin Heidelberg, Berlin, Heidelberg, (2011).