How To Infer The Informational Energy from Small Datasets

Angel Caţaron
Transilvania University of Braşov
Romania
Email: cataron@etc.unitbv.ro

Răzvan Andonie
Central Washington University
Ellensburg, USA
and
Transilvania University of Braşov
Romania
Email: andonie@cwu.edu

***Abstract*—Motivated by the problems in machine learning, we introduce a novel non-parametric estimator of Onicescu's informational energy. Our method is based on the k-th nearest neighbor distances between the n sample points, where k is a fixed positive integer. For some standard distributions, we investigate the performance of the estimator for small datasets.**

I. Introduction

Machine learning techniques based on inference are very much influenced by the size of the training set. When it comes to small training sets, the performance may not be so good, or the learning task can even not be accomplished. Small dataset conditions exist in many applications, such as disease diagnosis, fault diagnosis or deficiency detection in biology and biotechnology, mechanics, flexible manufacturing system scheduling, drug design, and short-term load forecasting (an activity conducted on a daily basis by electrical utilities). Several computational intelligence techniques have been proposed to overcome the limits of learning from small datasets. An overview of these techniques may be found in [1].

Before discussing the difficulties of inferring from small, or non-representative, training sets, we need to define formally what we understand by "small dataset". In many multivariable classification or regression (e.g., estimation or forecasting) problems we have a training set $T_p = (x_i, t_i)$ of p pairs of input/output vector $\mathbf{x} \in \Re^{\mathbf{n}}$ and scalar target t, and the unfortunate circumstance that T_p is small. The VC (Vapnik-Chervonenkis) dimension is a measure of the capacity of a classificator, defined as the cardinality of the largest set of points that the algorithm can shatter. According to Vapnik: "For estimating functions with VC dimension h, we consider the size p of data to be small if the ratio p/h is small (say $p/h < 20$)" [2].

The main reason why small datasets cannot provide enough information is that there exist gaps between samples, even the domain of samples cannot be ensured. For instance, in case of a small training set, even a simple neural network can have a complexity (e.g., number of connections/parameters) that is comparable to, or exceeds, the training size p. In such a case, we may expect to fit T_p very well. However, we can also expect poor generalization to new data identically distributed as the data in T_p. In effect, the VC dimension is too large relative to the size of the training set.

A completely different definition for "small" sets comes from algorithmic information theory. The Kolmogorov complexity of an object such as a string is a measure of the computational resources needed to specify the object. More formally, the complexity of a string is the length of the string's shortest description in some fixed universal description language. It can be shown that the Kolmogorov complexity of any string cannot be too much larger than the length of the string itself. A string is considered to be "random" if the length of the shortest problem that generates the string is the same as that of the string itself. Strings whose Kolmogorov complexity is small relative to the string's size are considered to have small information content [3]. Kolmogorov's complexity has been studied in the context of inductive inference [4], [5]. It is an open problem how to relate the Kolmogorov complexity of a training set and the generalization capability of the inferred neural network.

We will use a simplified definition: A training set is small if p and n are comparable.

There is no universally optimal solution to the problem of inferring from small datasets. Several techniques have been proposed [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]: generate artificial training samples, feature selection, and parameter fine-tuning of the inferred model. A special learning method designed for small training sets is the Central Location Tracking method [14], [15]. This algorithm attempts to explore the predictive information through the generation of trend value of each datum. The choice of specific technique is domain dependent.

Inference is based on a strong assumption: using a *representative* training set of samples to infer a model. In this case, we select a subset of the population, perform a statistical analysis on this sample, and use these results as an approximation to the desired statistical characteristics of the population as a whole. The more representative the sample, the larger our confidence that the statistical results obtained by using this sample are indeed a good approximation to the desired population statistics. We gauge the representativeness of a sample by how well its statistical characteristics reflect

978-1-4673-1653-8/12/$31.00 ©2012 IEEE

the statistical characteristics of the entire population. Many standard techniques may be used to select a representative sample set [16]. However, if we do not use expert knowledge, selecting the most representative training set from a given dataset was proved to be computationally difficult (NP-hard) [17]. The problem is actually more difficult, since in most applications the complete dataset is unknown or too large to be analyzed. Therefore, we have to rely on a more or less representative training set.

Another problem may arise from the training process itself. Especially in cases where learning was performed too long or where training the training samples are rare, the inferred model may adjust to very specific random features of the training data, that have no causal relation to the target function. In this process of *overfitting*, the performance on the training examples still increases while the performance on unseen data becomes worse (the generalization performance is poor). Beside preventing overfitting, a major question is how to detect it [1].

Many machine learning algorithms are based on information theory. A critical aspect of these approaches is how well an information theory measure is estimated from the available training set. This relates to a fundamental concept in statistics: probability density estimation. *Density estimation* is the construction of an estimate of the density function from the observed data [18]. We will refer here to *nonparametric estimation*, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density f, the data will be allowed to speak for themselves in determining the estimate of f more than would be the case if f were constrained to fall in a given parametric family. A common measure used in machine leaning is mutual information (MI). Several methods were proposed for density and MI estimation [19], [20], [21]. Nonparametric density estimators are histogram based estimator, adaptive partitioning of the XY plane, kernel density estimator (KDE), B-Spline estimator, k-th nearest neighbor (kNN) estimator and wavelet density estimator (WDE). Estimating MI techniques include histogram based, adaptive partitioning, spline, kernel density and kNN [20].

Estimating entropy and MI from small datasets is known to be a non-trivial task [22]. Naïve estimations (which attempt to construct a histogram where every point is the center of a sampling interval) are plagued with both systematic (bias) and statistical errors. An ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate tradeoff between the two types of errors [22].

MI is generally based on the classical Shannon type MI. However, it is computationally attractive to use one of its generalized forms: the Rényi divergence measure, which uses Rényi‘s quadratic entropy. The reason is that, as proved by Principe *et al.*, Rényi‘s quadratic entropy (and Rényi’s divergence measure) can be estimated from a set of samples using Parzen’s windows approach [23]. The MI and Rényi’s divergence measure are equivalent, but only in the limit $\alpha = 1$, where α is the order of Rényi‘s divergence measure [23]. The Parzen windows estimate cannot be performed on Shannon’s type MI.

In previous work, we have introduced a series of computational intelligence tools (classifiers and feature weighting / ranking techniques) based on an Onicescu’s informational energy (IE) and an unilateral dependency measure [24], [25], [26], [27], [28]. This measure proved to be an efficient alternative to the MI and we approximated it using the Parzen windows approach.

Our focus now is very different. The question is how to approximate (from a small dataset) the IE. Our contribution is a novel non-parametric biased approximation scheme, based on the kNN approach. First, we will review (Section II-A) the properties of IE and the kNN method. Section III introduces our theoretical result - an approximation method for the IE. After the experimental results exposed in Section IV, we will conclude with final remarks and some open problems (Section V).

II. Background

A. Onicescu’s Informational Energy

There are two strategies one can adopt when studying the relationship between two interacting systems: the first is to measure their interdependence thought as a mutual attribute and the second is to measure how much one system depends on the other. When the two systems are random variables, the most frequently used measures are based on information theory.

The MI is an example of the first strategy. Since it is a symmetric function, it measures simultaneously the dependence of one random variable by the other and vice versa. The second strategy can be illustrated by a *unilateral* measure which is not necessarily a symmetric function: such a measure was defined by Andonie *et al.* [29]. This unilateral measure is based on Onicescu’s IE [30]. We will introduce some of the IE basic concepts in this section.

Generally, information measures refer to uncertainty. Since Shannon defined his probabilistic information measure in 1948, many other authors, with Rényi, Daroczy, Bongard, Arimoto, and Guiaşu among them, have introduced new measures of information. The MI, $I(Y,X) = H(Y) - H(Y|X)$, measures the dependence between two random variables X and Y using Shannon’s entropy. In feature selection algorithms, the MI can be used for evaluating the ”information content” of each individual feature with regard to the output class. The feature selection method is searching for a subset of relevant features from an initial set of available features. The subset should maximize MI.

Information measures can also refer to certainty. Probability can be considered as a measure of certainty. More general, any monotonically growing and continuous function of a given probability can be considered as a measure of certainty.

Onicescu's IE was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration.

For a discrete random variable X with probabilities p_k, the IE was defined in [30] as:

$$IE(X) = \sum_{k=1}^{n} p_k^2. \tag{1}$$

For a continuous random variable Y the IE was defined in [31]:

$$IE(Y) = \int_{-\infty}^{+\infty} f^2(\mathbf{y}) d\mathbf{y}, \tag{2}$$

where $f(\mathbf{y})$ is the probability density function of the random variable.

In order to study the interaction between two random variables X and Y, the following measure of unilateral dependency was defined by Andonie *et al.* [29]:

$$o(Y, X) = IE(Y|X) - IE(Y)$$

This measure quantifies the unilateral dependence characterizing Y with respect to X and corresponds to the amount of information detained by X about Y. There is an obvious analogy between $o(Y, X)$ and the MI, since both measure the same phenomenon. However, the MI is a symmetric, not a unilateral measure.

Rather than approximating $o(Y, X)$ as we did in our previous studies, we will approximate now directly the IE from the available dataset.

B. The nearest neighbor method

The nearest neighbour class of estimators represents an attempt to adapt the amount of smoothing to the "local" density of data. The degree of smoothing is controlled by an integer k, chosen to be considerably smaller than the sample size; typically $k \approx n^{1/2}$. Define the distance $d(x, y)$ between two points on the line to be $|x - y|$ in the usual way, and for each t define $d_1(t) \leq d_2(t) \leq \ldots \leq d_n(t)$ to be the distances, arranged in ascending order, from t to the points of the sample.

The kNN density estimate $f(t)$ is then defined by [18]:

$$\hat{f}(t) = \frac{k}{2nd_k(t)} \tag{3}$$

The kNN was used for non-parametrical estimate of the entropy based on the k-th nearest neighbor distance between n points in a sample, where k is a fixed parameter and $k \leq n - 1$. Based on the first nearest neighbor distances, Leonenko *et al.* [32] introduced an asymptotic unbiased and consistent estimator H_n of the entropy $H(f)$ in a multidimensional space. When the sample points are very close one to each other, small fluctuations in their distances produce high fluctuations of H_n. In order to overcome this problem, Singh *et al.* [33] defined an entropy estimator based on the k-th nearest neighbor distances. A kNN estimate of the Kullback-Leibler divergence was obtained by Wang *et al.* in [34]. A mean of several kNN estimators corresponding to different values of k was used by Faivishevsky *et al.* in [35] for developing the smooth estimator MeanNN of differential entropy, mutual information and divergence.

We are ready now to introduce our kNN method for IE approximation.

III. Estimation of the Informational Energy

Our goal is to estimate (2) from a random sample X_1, X_2, ..., X_n of n d-dimensional realizations of a distribution with the unknown probability density $f(x)$. The IE is the average of $f(x)$, therefore we have to estimate $f(x)$. The n realizations from our samples have the same probability $\frac{1}{n}$. A convenient estimator of the IE is:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(X_i). \tag{4}$$

We will determine first the probability density $P_{ik}(\epsilon)$ of the distance $R_{i,k,n}$ between a fixed point X_i and its k-th nearest neighbor from the remaining $n-1$ points. Probability $P_{ik}(\epsilon)d\epsilon$ of the k-th nearest neighbor to be within distance $R_{i,k,n} \in [\epsilon, \epsilon + d\epsilon]$ from X_i, $k-1$ points at a smaller distance and $n-k-1$ at a larger distance can be expressed in terms of the trinomial formula [35]:

$$P_{ik}(\epsilon)d\epsilon = \frac{(n-1)!}{1!(k-1)!(n-k-1)!} dp_i(\epsilon) p_i^{k-1} (1-p_i)^{n-k-1},$$

where $p_i(\epsilon) = \int_{\|x - X_i\| < \epsilon} f(x)dx$ is the mass of the ϵ-ball centered at X_i and $\int P_{ik}(\epsilon)d\epsilon = 1$.

We can express the expected value of the $p_i(\epsilon)$ using the probability mass function of the trinomial distribution:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = \int_0^{\infty} P_{ik}(\epsilon) p_i(\epsilon) d\epsilon =$$

$$= k \binom{n-1}{k} \int_0^1 p^{k-1} (1-p)^{n-k-1} p dp =$$

$$= k \binom{n-1}{k} \int_0^1 p^{(k+1)-1} (1-p)^{(n-k)-1} dp.$$

This equality can be reformulated using the *Beta* function:

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

We obtain:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = k \binom{n-1}{k} \frac{\Gamma(k+1)\Gamma(n-k)}{\Gamma(n+1)} =$$

$$= k \frac{(n-1)!}{(n-k-1)!k!} \frac{k!(n-k-1)!}{n!},$$

which can be rewritten as:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = \frac{k}{n}. \tag{5}$$

On the other hand, assuming that $f(x)$ is almost constant in the entire ϵ-ball around X_i [35], we have:

$$p_i(\epsilon) \approx V_1 R_{i,k,n}^d f(X_i),$$

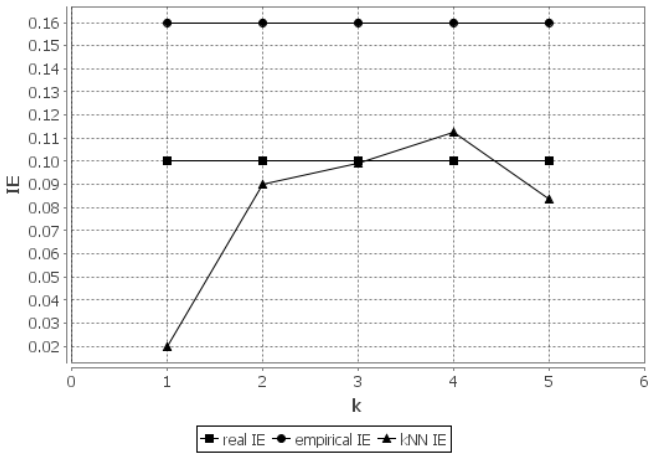

Fig. 1. The informational energy of an uni-dimensional sample with values obtained from a discrete uniform distribution.

where we denote the volume of the ball of radius r in a d-dimensional space by:

$$V_r = \frac{\pi^{d/2} r^d}{\Gamma(d/2+1)} = V_1 r^d,$$

V_1 is the volume of the unit ball and $R_{i,k,n}$ is the Euclidean distance between the reference point X_i and its k-th nearest neighbor. Thus, $V_1 R_{i,k,n}^d$ is the volume of the ball of radius $R_{i,k,n}$. We obtain the expected value of $p_i(\epsilon)$:

$$E(p_i(\epsilon)) = E(V_1 R_{i,k,n}^d f(X_i)) = V_1 R_{i,k,n}^d \hat{f}(X_i). \quad (6)$$

Both equations (5) and (6) estimate $E(p_i(\epsilon))$. We have:

$$V_1 R_{i,k,n}^d \hat{f}(X_i) = \frac{k}{n},$$

$$\hat{f}(X_i) = \frac{k}{n V_1 R_{i,k,n}^d}, i = 1...n.$$

This is the estimate of the probability density function. By substituting $\hat{f}(X_i)$ in (4), we finally obtain the IE approximation:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{k}{n V_1 R_{i,k,n}^d}. \quad (7)$$

IV. Experiments

The IE can be easily computed if the data sample is extracted from known distributions. When the underlying distribution of data sample is unknown, the IE has to be estimated. The problem is even more difficult if the number of available points is small.

The tests presented below use data points generated with normal and uniform distribution, enabling us to find the real value of the IE. For normal distributions, we use uni and bi-dimensional data. In case of uniform distributions, we perform the tests on one and ten-dimensional data.

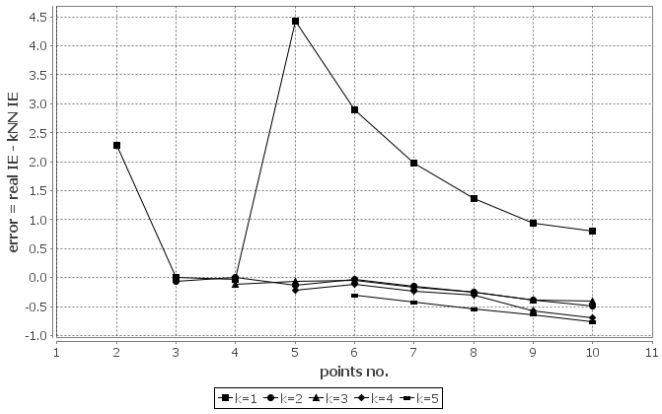

Fig. 2. The informational energy of an uni-dimensional sample with values obtained from a continuous normal distribution with mean 0 and variance 1.

A. A simple uni-dimensional example

To illustrate why the approximation of information measures from a limited number of data samples is non-trivial, we will start with a very simple example. Let us consider the following uni-dimensional data samples: $\{3, 4, 1, 2, 8, 10, 1, 3, 4, 9\}$ obtained from the discrete uniform distribution $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The "true" IE value can be directly obtained from definition (1):

$$IE_{true} = \sum_{i=1}^{10} \left(\frac{1}{10}\right)^2 = 0.1.$$

Assuming that the real distribution is unknown, which is the interesting case, we can use the relative frequencies in formula (1) to obtain the "empirical" IE. The relative frequencies are $\{2, 8, 9, 10\} : \frac{1}{10}$ and $\{1, 3, 4\} : \frac{2}{10}$. We obtain:

$$IE_{empirical} = 4 \cdot \left(\frac{1}{10}\right)^2 + 3 \cdot \left(\frac{2}{10}\right)^2 = 0.16.$$

The $IE_{empirical}$ is not a good estimate especially when the relative frequencies are far from the true probabilities. This is generally the case for small datasets and, in accordance to the central limit theorem, for an increasing number of samples, $IE_{empirical}$ converges probabilistically to IE_{true}.

The kNN estimator of the IE defined by (7) uses a fixed k for finding the k-order statistics of Euclidean distances between the fixed point X_i and all other $n-1$ points. To avoid the division by 0, we ignore the cases when $R_{i,k,n} = 0$. The kNN estimation of the information energy of our sample is depicted in Figure 1 for $k = 1...5$. The estimation error is high when $k = 1$, i.e. the *nearest neighbor* estimation, but the $\hat{IE}$ is close to IE_{true} for $k > 1$.

B. Uni-dimensional normal distribution

Let us consider the continuous uni-dimensional normal distribution with the mean 0 and variance 1. Figure 2 illustrates the difference between IE_{real} and $\hat{IE}$ in five experiments on small datasets with samples of 2 to 10 points ($k = 1, 2, \ldots, 5$). We omit the case when k is greater than the number of

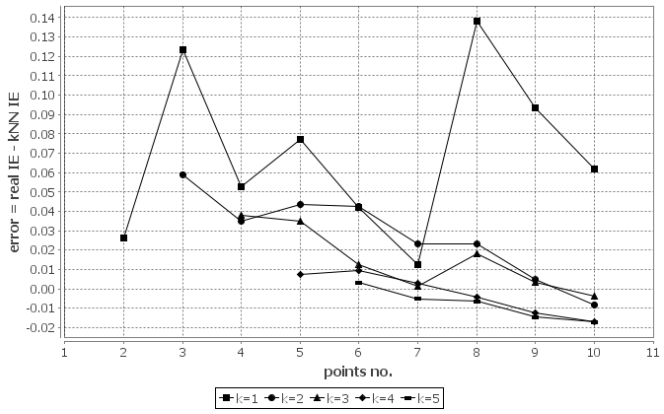

Fig. 3. The informational energy of an uni-dimensional sample with values obtained from a continuous normal distribution with mean 0 and variance 1.

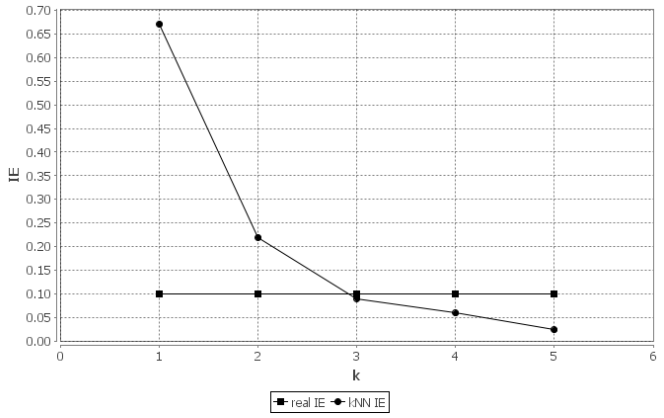

Fig. 4. The informational energy of a ten-dimensional sample with values obtained from a discrete uniform distribution.

samples. The *1-nearest neighbor* estimator is unstable, showing large variations around IE_{real}. Increasing the value of k results in an approximation with a slight biasing tendency. This bias becomes more visible with an increasing number of samples. For small datasets, this approximation bias is less relevant.

C. Bi-dimensional normal distribution

We consider now the bi-dimensional normal distribution with mean:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and covariance matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The conditions of this experiment are similar to the bi-dimensional normal distribution. On the small datasets of $2, \ldots, 10$ points we applied the kNN estimation with $k = 1, 2, \ldots, 5$. The error between IE_{real} and $\hat{IE}$ stabilizes for larger values of k, as presented in Figure 3.

D. Ten-dimensional uniform distribution

Finally, we will use the kNN IE estimator for a case when dimensionality equals the number of data samples. We use the discrete uniform distribution with ten values per dimension. Figure 4 depicts the obtained approximation (IE_{real} versus $\hat{IE}$).

V. Conclusion and Open Problems

We have introduced a novel non-parametric kNN approximation method for computing the IE from data samples. According to our experiments, the method proves to be more accurate for small datasets because the approximation bias is less influential in this case.

It is possible to use this kNN approach to approximate the dependency measure $o(Y, X)$. In this case, it would be interesting to compare this approximation with the one obtained by Parzen windows in our previous work.

We are are presently studying the asymptotic behaviour of the mean and variance of this approximator in order to prove its consistency. In the second stage, we plan to obtain an unbiased version of it. We also plan to apply our IE estimator to several machine learning techniques were the IE and the unilateral dependency can be used, especially whenever only small datasets are available: feature extraction and ranking, classification, prediction.

References

[1] R. Andonie, "Extreme data mining: Inference from small datasets," *International Journal of Computers, Communications and Control*, vol. 5, pp. 280–291, 2010.

[2] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 2000.

[3] J. L. Balcázar and R. V. Book, "Sets with small generalized Kolmogorov complexity," *Acta Inf.*, vol. 23, no. 6, pp. 679–688, 1986.

[4] A. Ambainis, "Application of Kolmogorov complexity to inductive inference with limited memory," in *ALT '95: Proceedings of the 6th International Conference on Algorithmic Learning Theory*. London, UK: Springer-Verlag, 1995, pp. 313–318.

[5] A. Ambainis, K. Apsitis, C. Calude, R. Freivalds, M. Karpinski, T. Larfeldt, I. Sala, and J. Smotrovs, "Effects of Kolmogorov complexity present in inductive inference as well," in *ALT '97: Proceedings of the 8th International Conference on Algorithmic Learning Theory*. London, UK: Springer-Verlag, 1997, pp. 244–259.

[6] J.-L. Yuan and T. Fine, "Neural-network design for small training sets of high dimension," *IEEE Tnansactions on Neural Networks*, vol. 9, pp. 266–280, 1998.

[7] J.-L. Yuan, "Bootstrapping nonparametric feature selection algorithms for mining small data sets," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1999, pp. 2526 – 2529.

[8] C. Huang and C. Moraga, "A diffusion-neural-network for learning from small samples," *International Journal of Approximate Reasoning*, vol. 35, pp. 137–161, 2004.

[9] R. Mao, H. Zhu, L. Zhang, and A. Chen, "A new method to assist small data set neural network learning," in *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA06)*, 2006, pp. 17–22.

[10] D.-C. Li, C.-S. Wu, T. T.-I., and L. Y.-S., "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," *Computers and Operations Research*, vol. 34, pp. 966–982, 2007.

[11] D.-C. Li, C.-W. Yeh, T.-I. Tsai, Y.-H. Fang, and S. Hu, "Acquiring knowledge with limited experience," *Expert Systems*, vol. 24, pp. 162–170, 2007.

[12] D.-C. Li, C.-S. Wu, T.-I. Tsai, and F. M. Chang, "Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge," *Comput. Oper. Res.*, vol. 33, no. 6, pp. 1857–1869, 2006.

[13] T.-I. Tsai and D.-C. Li, "Approximate modeling for high order non-linear functions using small sample sets," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 564–569, 2008.

[14] D.-C. Li and C.-W. Yeh, "A non-parametric learning algorithm for small manufacturing data sets," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 391–398, 2008.

[15] D.-C. Li and C.-W. Liu, "A neural network weight determination model designed uniquely for small data set learning," *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9853–9858, 2009.

[16] H. Lohr, *Sampling: Design and Analysis*. Duxbury Press, 1999.

[17] J. Gamez, F. Modave, and O. Kosheleva, "Selecting the most representative sample is NP-hard: Need for expert (fuzzy) knowledge," in *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, June 2008, pp. 1069–1074.

[18] B. Silverman, *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1986.

[19] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1253, June 2003.

[20] J. Walters-Williams and Y. Li, "Estimation of mutual information: A survey," in *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 389–396.

[21] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, pp. 1–16, Jun 2004.

[22] J. A. Bonachela, H. Hinrichsen, and M. A. Munoz, "Entropy estimates of small data sets," *MATH.THEOR.*, vol. 41, p. 202001, 2008.

[23] J. C. Principe, D. Xu, and J. W. F. III., "Information-theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed., Wiley, New York, 2000.

[24] R. Andonie and A. Cațaron, "An informational energy LVQ approach for feature ranking," in *European Symposium on Artificial Neural Networks 2004, pages In d-side publications*, 2004, pp. 471–476.

[25] A. Cațaron and R. Andonie, "Energy generalized lvq with relevance factors," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 2, july 2004, pp. 1421 – 1426 vol.2.

[26] ——, "Informational energy kernel for lvq," in *Proceedings of the 15th international conference on Artificial neural networks: formal models and their applications - Volume Part II*, ser. ICANN'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 601–606.

[27] R. Andonie, "How to learn from small training sets," Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno-Lugano, Switzerland, invited talk, September 2009.

[28] A. Cațaron and R. Andonie, "Energy supervised relevance neural gas for feature ranking," *Neural Processing Letters*, vol. 32, no. 1, pp. 59–73, 2010.

[29] R. Andonie and F. Petrescu., "Interacting systems and informational energy," *Foundation of Control Engineering*, no. 11, pp. 53–59, 1986.

[30] O. Onicescu, "Theorie de l'information. energie informationelle," *C. R. Acad. Sci. Paris, Ser. A–B*, no. 263, pp. 841–842, 1966.

[31] S. Guiasu, *Information theory with applications*. McGraw Hill New York, 1977.

[32] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Probl. Peredachi Inf.*, vol. 23, no. 2, pp. 9–16, 1987.

[33] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neightboor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, pp. 301–321, 2003.

[34] Q. Wang, S. R. Kulkarni, and S. Verdu, "A nearest-neighbor approach to estimating divergence between continuous random vectors," in *Proc. of the IEEE International Symposium on Information Theory*, Seattle, WA, 2006.

[35] L. Faivishevsky and J. Goldberger, "Ica based on a smooth estimation of the differential entropy," in *NIPS*, 2008.