The Uncertainty Principle of Cross-Validation

Mark Last, *Member, IEEE*

***Abstract* - Data miners have often to deal with data sets of limited size due to economic, timing and other constraints. Usually their task is two-fold: to induce the most accurate model from a given dataset and to estimate the model's accuracy on future (unseen) examples. Cross-validation is the most common approach to estimating the true accuracy of a given model and it is based on splitting the available sample between a training set and a validation set. The practical experience shows that any cross-validation method suffers from either an optimistic or a pessimistic bias in some domains. In this paper, we present a series of large-scale experiments on artificial and real-world datasets, where we study the relationship between the model's true accuracy and its cross-validation estimator. Two stable classification algorithms (ID3 and info-fuzzy network) are used for inducing each model. The results of our experiments have a striking resemblance to the well-known Heisenberg Uncertainty Principle: the more accurate is a model induced from a small amount of real-world data, the less reliable are the values of simultaneously measured cross-validation estimates. We suggest calling this phenomenon "the uncertainty principle of cross-validation".**

***Index Terms*— Cross-Validation, Accuracy Estimation, Model Selection, Classification, Info-Fuzzy Networks.**

I. INTRODUCTION

Though data mining is traditionally associated with extracting knowledge from *large amounts of data* [7], the real-world data miners have often to deal with much smaller data sets than they would prefer to have. This situation may arise, for instance, in the health care sector, where a hospital may be able to provide only a limited number of patient records, or in analysis of a new manufacturing process, where each record contains the results of a costly engineering experiment. In this paper, we mainly focus on *classification* problems of data mining. Given the set of available data and a classification algorithm of his/her choice, the data miner faces a two-fold task: to induce the most accurate model from a given dataset and to estimate the true model's accuracy on future (unseen) examples. For many users, the second task is no less important than the first one, since they would like to choose the *most accurate* model of *known accuracy* from a given set of alternative models.Obviously, this choice has to be made under financial, timing, and other constraints that limit the amount of available data.

The problem of selecting the most accurate classifier among several classifiers is known as the problem of *model selection* [10]. There are multiple ways to form a set of alternative models such as varying the complexity of a given hypothesis [8] or running different induction algorithms against the same data set [10]. The goal of the corresponding model selection techniques is finding the best complexity level or choosing the most accurate inducer. The inducer selection problem has been treated extensively in machine learning and data mining literature and we leave it beyond the scope of this paper. Our assumption is that as defined in [6] the best induction algorithm(s) have been chosen at the pre-DM stage of the knowledge discovery process based on performance statistics, user preferences, software costs, and other criteria. For the sake of simplicity, we also assume that the user cannot explicitly control the complexity of the induced model like in the case of [8]. Thus, our model selection problem is reduced to the following definition: *using the selected induction algorithm with a pre-specified set of parameters, induce the most accurate model from a given dataset*.

As indicated in the beginning of this section, the above problem of model selection is tightly coupled with the problem of *accuracy estimation*: we want to find the most accurate model, but we also want to know how accurate it is. This excludes the option of taking the entire available sample as the training set, since the *apparent (re-substitution) accuracy* is well known to be an overoptimistic estimator [7]. Having virtually unlimited CPU power, we can obtain an interval estimator of the classifier's accuracy by some method of *k-fold cross-validation* and then induce a single model from the entire sample. However cross-validation estimators (especially, the *leave-one-out* method) are known to suffer from high variance, resulting in relatively wide confidence intervals, and, as shown by Kohavi in [10], that variance can be partially reduced only at the cost of increasing the bias between the estimated and the true accuracy. Consequently, it may be reasonable to use the cross-validation procedure for separating the good inducers from the poor ones, but the same procedure is hardly helpful for comparing classification models of similar accuracy.

To induce both a classification model and an estimator of its true accuracy from the same data sample, we can apply a *two-fold cross validation* also

Manuscript received December 31, 2005. This work was supported in part by a research contract from the Israel Ministry of Defense and by the National Institute for Systems Test and Productivity at University of South Florida under the USA Space and Naval Warfare Systems Command Grant No. N00039-01-1-2248.

M. Last is with the Dept. of Information Systems Eng., Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (phone: +972-8-6461397; fax: +972-8-6477527; e-mail: mlast@bgu.ac.il).

called the *holdout* method [7]. In this method, the sample is randomly partitioned into a *training set* that is used to induce the model and a *test set* where we evaluate the model's accuracy. The underlying assumption is that the induction algorithm is semantically stable, i.e. the classifiers produced by different random partitions of the dataset are expected to make the same predictions for the same instances. Typically, 33% or 50% of the data are held out for testing the model, though no theoretical justification for these or other splits are known [2]. When we are interested in the mean accuracy of a given inducer rather than in the true accuracy of a specific model, this procedure can be repeated multiple times. This is called *holdout with random subsampling*. Since the classification performance of a typical inducer is a non-decreasing function of the number of training instances, the two-fold CV estimator, like its multi-fold counterparts, tend to be highly pessimistic [7]. Techniques for reaching a better balance between an optimistic and a pessimistic estimator include *stratified cross validation* [3] and the family of *bootstrap methods* [4] [5]. However, Kohavi has shown in [10] that none of these techniques, combined with a varying number of cross-validation folds, is guaranteed to provide unbiased estimations for any domain.

Kearns [9] has made an attempt to find the optimal value of the training-test split rather than using the standard 2:1 ratio. He has developed an analytical expression for calculating the optimal value of the test data fraction as a function of the sample size and the target concept complexity. Since in a real-world problem the complexity of the underlying concept may be unknown in advance, Kearns has performed a series of controlled experiments with data produced by *noiseless* models of known complexity. The case studies were characterized by the power law behavior of the learning curve. His main conclusion is that as long as the complexity is small compared to the sample size, the generalization performance of two-fold cross validation is rather insensitive to the choice of the split ratio. The second conclusion implies that choosing the same fraction of testing records (about 0.5) will nearly minimize the cross validation error bound for a wide range of target functions. The results of [9] provide no clue as to the value, or even the existence of an optimal split ratio for real-world data governed by noisy concepts and for inducers that may deviate from the power law behavior.

This paper is organized as follows. Section II covers a series of large-scale experiments on artificial and real-world datasets, where we study the relationship between the model's true accuracy and its cross-validation estimator as a function of the training/test split ratio. In Section III, we try to draw a parallel between our observations and the well-known Heisenberg Uncertainty Principle of Quantum Mechanics. Implications for the practical process of knowledge discovery and open research topics are suggested in Section IV.

II. EMPIRICAL RESULTS: ACCURACY AS A MOVING TARGET

A. Experimental Settings

As indicated in Section I above, this paper considers only the task of selecting a single most accurate model of known accuracy rather than selecting the most accurate inducer for a given dataset. Consequently, all our experiments will focus on the performance of a two-fold (rather than 5 or 10-fold) cross-validation as a function of the training-test ratio. A necessary condition for *k*-fold cross-validation to be an accurate estimator of predictive accuracy is semantic stability of the induction algorithm. This implies that if an algorithm is stable for a given dataset, the variance of the cross-validation estimates should be nearly independent of the number of folds [10]. For our large-scale experiments, we have chosen ID3 [15], which is a common and relatively stable decision-tree algorithm and the Info-Fuzzy Network (IFN) classifier, which was shown in [13] to produce more compact and stable decision-tree models than other decision-tree algorithms (including C4.5), while preserving nearly the same level of predictive accuracy. Info-fuzzy network is an oblivious read-once decision graph built by a top-down information-theoretic algorithm, which uses the likelihood-ratio test as a pre-pruning criterion. More details on IFN can be found in [12] and [14].

We have used the same data sets from a wide variety of real-world domains that were used in Kohavi's cross-validation experiments [10]. All data sets including the *no information* set of artificially generated random data were downloaded from the MLC++ web site [11]. Most of them originate from the UCI Machine Learning Repository [1]. We have also added a *perfect information dataset*, which includes 3,000 records of non-random noiseless data having the same values of input attributes as Kohavi's "no information" set. The underlying model we used to determine the values of the target attribute in the perfect information dataset is shown in Fig. 1. As one can see, only three features (out of 20 available inputs) are relevant for predicting the target value, while the probability of class = 0, under the assumption that all input attributes are random, is 0.625.

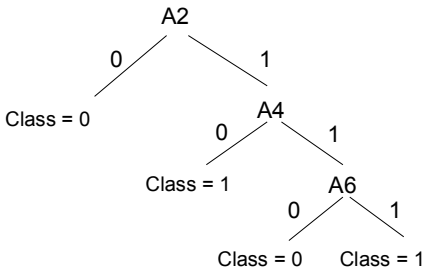

Fig. 1 Perfect Information Dataset

The datasets were evaluated with a sample size around the point where the learning curves of both algorithms flattened out. The learning curves of ID3 and IFN algorithms are shown for all datasets in the case study in Fig. 2 and Fig. 3 respectively as a function of the fraction of the initial dataset that was used for training the algorithm (0% - 50%). The predictive accuracy was measured on the remaining test instances. The value of Training Set Size = 0% refers to the accuracy of the default (majority) rule on the entire dataset. Most datasets (including the *perfect information* dataset) seem to reach nearly the maximum accuracy level for only 5% of the available data.

The "true" accuracy of each inducer for a given partition of every dataset was estimated using 100 runs of the holdout method. In Table 1, we show the sample percentage and the estimated accuracy of each classifier.

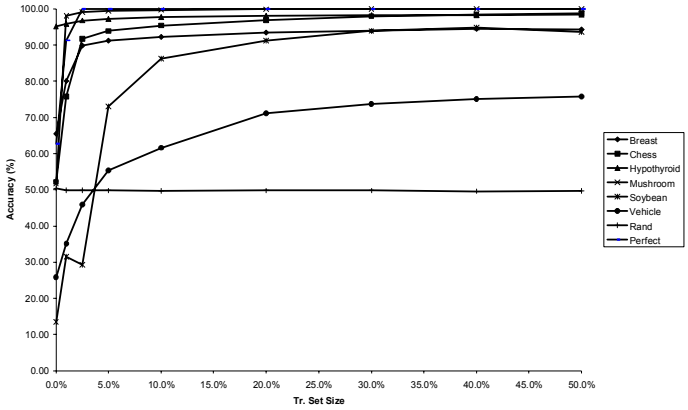

Fig. 2 ID3 Learning Curves

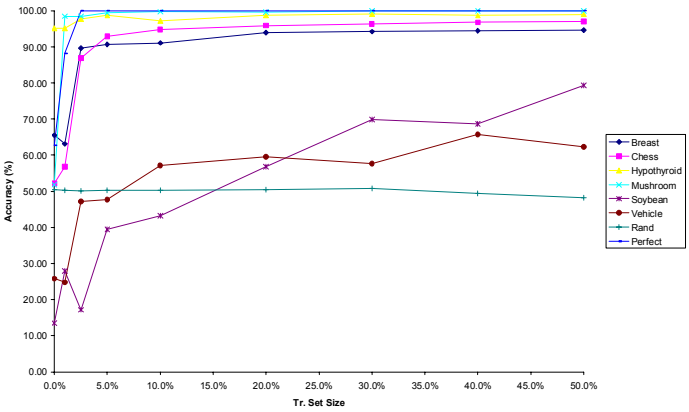

Fig. 3 IFN Learning Curves

TABLE 1
EXPECTED ACCURACY ESTIMATES FOR ID3 AND IFN

			ID3		IFN	
Dataset	Number of Attributes	Total Size	Sample Percentage	C.I.	Sample Percentage	C.I.
Breast Cancer	10	699	5%	91.2+-0.55	5%	90.46+-0.5
Chess	36	3196	5%	93.9+-0.46	5%	93.26+-0.53
Hypothyroid	25	3163	5%	97.2+-0.32	5%	97.2+-0.13
Mushroom	22	8124	5%	99.5+-0.13	5%	99.33+-0.07
Soybean large	35	683	30%	93.9+-0.46	50%	71.24+-1.19
Vehicle	18	846	20%	71.1+-0.88	10%	56.82+-1.01
Rand	20	3000	5%	49.8+-0.97	5%	49.91+-0.09
Perfect	20	3000	5%	100+-0	5%	100+-0

In our experiments with all datasets, we have varied the test fraction of the selected sample size between 10% and 90%, while using the instances not included in the sample for estimating the true accuracy of the induced model. This was repeated 100 times for each value of the test fraction. The quality of cross-validation results was measured using the following criteria: *average true accuracy* (the average accuracy of 100 models measured on the same validation instances not included in the sample) and *correlation coefficient* between the cross-validation estimate and the true accuracy of each model (measured over 100 value-pairs of estimated and true accuracy). The latter criterion is particularly important for evaluating reliability of cross-validation estimates, since to choose the best model, we are interested in predicting the true accuracy of each alternative model based on its cross-validation estimate (see Section I above).

A. Analysis of Results

Figs. 4 – 11 present the trade-off between the true error and the correlation coefficient for each real-world and artificial data set in our study. Ideally, we would like to find on each plot a point, which is relatively close to (0.00, 1.00), where the error rate is zero and the correlation coefficient is one. Unfortunately, this occurs only in the Perfect dataset (Fig. 11) which contains noiseless data. In all other datasets, except Rand and Hypothyroid where the error rate is nearly constant, any significant improvement in correlation can be reached only at the cost of increased error rate. We suggest to call this empirical phenomenon the Uncertainty Principle of Cross-Validation: *The more accurate is a model induced from a small amount of real-world data, the less reliable are the values of simultaneously measured cross-validation estimates*.

Though this uncertainty phenomenon deserves a careful study, we may try to suggest here a partial explanation as follows. More accurate models induced from larger training sets are more complex, which implies that they are more sensitive to small changes in the training data. Such changes may lead to either generation of occasional spurious patterns or, simultaneously a loss of some valid patterns that present in the data. Each change like this may affect the model's structure and its predictive accuracy. On the other hand, when the training set is small, the induced model will be simpler and more stable. Now we proceed with drawing some basic parallels between the Uncertainty Principle of Cross-Validation and the Heisenberg Uncertainty Principle of Quantum Mechanics.

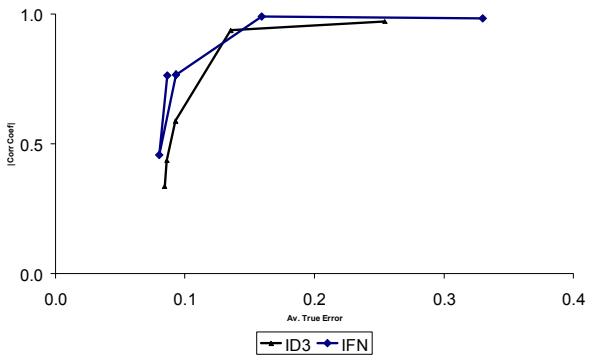

Fig. 4 Breast Dataset: Accuracy vs. Correlation

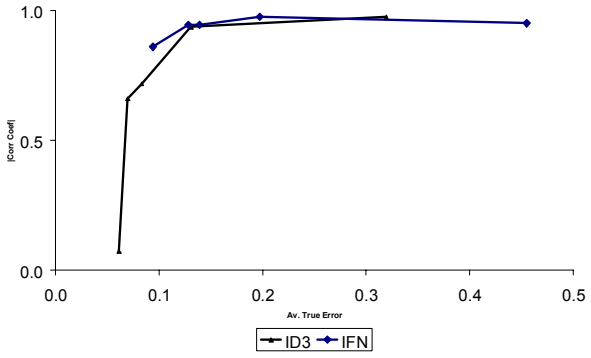

Fig. 5 Chess Dataset: Accuracy vs. Correlation

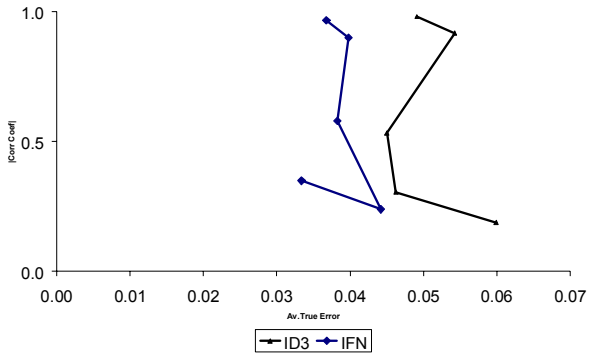

Fig. 6 Hypothyroid Dataset: Accuracy vs. Correlation

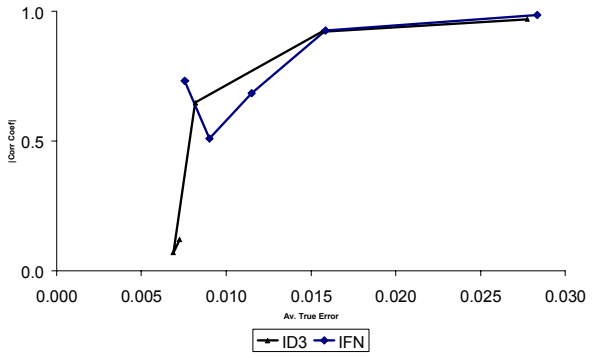

Fig. 7 Mushroom Dataset: Accuracy vs. Correlation

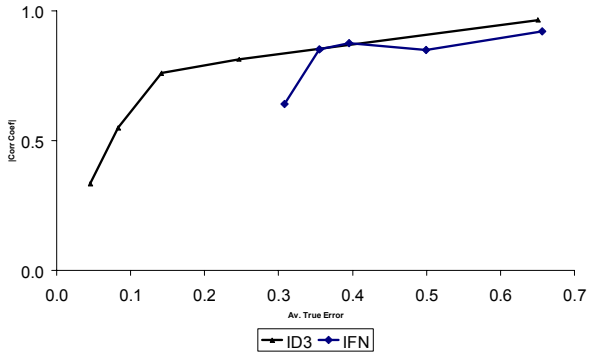

Fig. 8 Soybean Dataset: Accuracy vs. Correlation

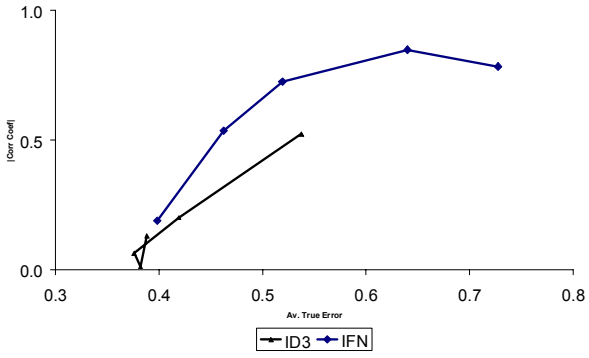

Fig. 9 Vehicle Dataset: Accuracy vs. Correlation

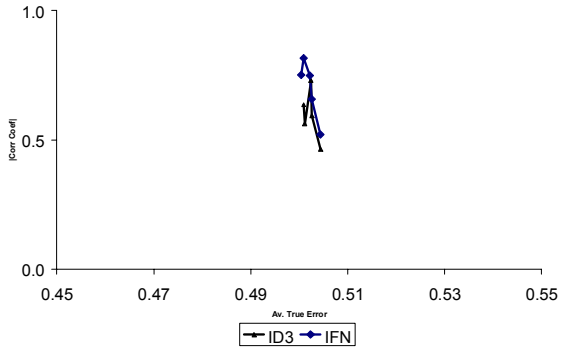

Fig. 10 Rand Dataset: Accuracy vs. Correlation

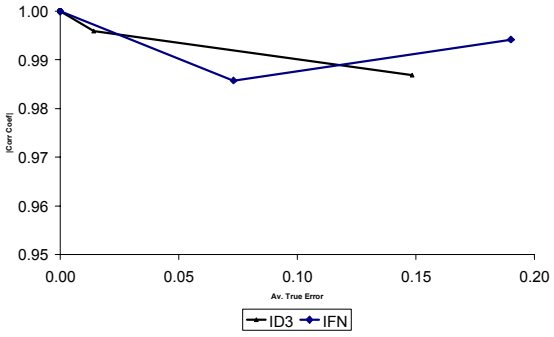

Fig. 11 Perfect Dataset: Accuracy vs. Correlation

III. A BRIEF DISCUSSION OF THE UNCERTAINTY PRINCIPLE

The Heisenberg Uncertainty Principle [16] says that we cannot know the exact values of two physical quantities that describe an *atomic* system. Examples include simultaneous measurement of the position and momentum of a particle or of energy and time. A common interpretation of this principle is as follows: we cannot determine the exact value of one quantity without losing all information on the other quantity. In the intermediate cases, there is a known relationship (based on the Planck's constant) between the uncertainties of the two

simultaneously measured values. As indicated above, all this is true only for systems of *atomic* order.

Though the exact nature of cross-validation uncertainty may be quite different from the laws of Quantum Mechanics, we would like to point out some obvious similarities of these two phenomena:

The Uncertainty Principle of Cross-Validation applies only to samples of small ("atomic") size. Having an unlimited amount of data, we should have no difficulty with inducing an accurate model of *known* accuracy.

The most accurate model can only be induced from *all* available data. However, in this case we cannot estimate the true accuracy of the obtained model.

In another extreme case, we may base our prediction on a prior belief rather than any training data. Estimation of such belief's accuracy on a validation set of maximum size would be highly reliable.

As shown by the empirical results of Section II, there appears to be an inverse relationship between the true accuracy and the capability to estimate it reliably in most datasets. Obviously, these results are not sufficient for finding the exact form of this relationship or for claiming that it can be found in *any* dataset.

IV. CONCLUSIONS

In this paper, we have explored on large-scale case studies the problems associated with reliable estimations of classifier accuracy using cross-validation techniques and finite-size data samples. The results of our experiments have a striking resemblance to the well-known Heisenberg Uncertainty Principle: the more accurate is a model induced from a small amount of real-world data, the less reliable are the values of simultaneously measured cross-validation estimates. We suggest to call this phenomenon "the uncertainty principle of cross-validation". In our view, this important limitation of cross-validation techniques should be taken into consideration when planning the process of knowledge discovery in databases of limited size.

More experimentation is needed to better understand the universal relationship (if it exists) between uncertainties associated with the cross-validation process. The effect of algorithms' stability on the reliability of cross-validation estimates should also be studied in more detail.

REFERENCES

[1] Blake, C.L. & Merz, C.J. *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[2] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, P.J. *Classification and Regression Trees*. Wadsworth, 1984.

[3] Diamantidis N.A., Karlis D., and Giakoumakis, E.A. Unsupervised Stratification of Cross-Validation for Accuracy Estimation. *Artificial Intelligence*, 116 (2000), 1–16.

[4] Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78, 382 (Jun. 1983), 316-331.

[5] Efron, B., and Tibshirani, R. Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*, 92, 438 (Jun. 1997), 548-560.

[6] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining,* Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds., AAAI/MIT Press, 1996, 1-36.

[7] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[8] Kearns, M., Mansour, Y., Ng, A. Y., and Ron, D. An Experimental and Theoretical Comparison of Model Selection Methods. In *Proceedings of the seventh workshop on Computational Learning Theory(COLT)* , ACM Press, 1995, 21-30.

[9] Kearns, M. J. A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, With Consequences for the Training-Test Split. *Neural Computation*, 9, 5 (July 1997), 1143 - 1161.

[10] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Montréal, Québec, Canada, August 20-25, 1995). Morgan Kaufmann, 1995, 1137-1145.

[11] Kohavi, R. MLC++: A Library of C++ Classes for Supervised Machine Learning. *Silicon Graphics, Inc* (2004). [http://www.sgi.com/tech/mlc/].

[12] Last M., and Maimon, O. A Compact and Accurate Model for Classification. *IEEE Transactions on Knowledge and Data Engineering*, 16, 2 (Feb. 2004), 203-215.

[13] Last, M., Maimon, O., and Minkov, E. Improving Stability of Decision Trees. *International Journal of Pattern Recognition and Artificial Intelligence*, 16, 2 (2002), 145-159.

[14] Maimon O., and Last, M. *Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology*. Kluwer Academic Publishers, Massive Computing, Boston, December 2000.

[15] Quinlan, J.R. Induction of Decision Trees. *Machine Learning*, 1, 1 (1986), 81-106.

[16] Schiff, L.I. *Quantum Mechanics*. McGraw-Hill, New York, 1955