Comparison of Alternative Criteria for the Evaluation of Machine Learning in the Medical Diagnosis of Stroke

Ioannis Nomikos, Georgios Dounias, Konstantinos Vemmos*
Department of Production Engineering and Management
Technical University of Crete
University Campus, Kounoupidiana, 73100, Chania, Greece
Phone: +30-821-37323, Fax: +30-821-69410
email: inomikos@hotmail.com
*Unit of Acute Stroke, Therapeutic Clinic, "Alexandra" General Hospital
Athens, Greece

ABSTRACT: Decision tree induction is based on selection measures responsible for assigning importance degrees to the features involved in the classification process. In this paper we will present some alternative criteria and their variations for the purpose of comparing their effectiveness in the medical diagnosis of stroke. To avoid complicating the presentation with mathematical details we will focus on analyzing the results of the various measures.

KEYWORDS: Machine learning, decision tree induction, selection measures, stroke diagnosis

INTRODUCTION

Machine learning comprises of methods used to specify the rules underlying certain processes we deal with in our vital environment and to simulate an expert's analytical case approach. It uses a variety of methods to achieve this, one of which is a tree like structure establishing the logical routes an expert would follow to reach a decision and classify accordingly the case in question. Decision trees are common classifiers using induction algorithms in order to learn from an amount of given cases and create the appropriate conjunctions that will correspond to the process they aim to simulate. Each case is being described by a number of features, which the induction algorithm has to judge in terms of usefulness and properly place within the body of the decision tree. On the schema of this classification process the features are presented as inner nodes and are connected to the leaves. Each leaf corresponds to a class, though more than one leaves may belong to the same class. The leaves and inner nodes are being constructed during the decision tree's learning phase. To this purpose one uses selection measures to assign a specific value to each feature according to its contribution to the evaluation of the classes during the learning process. As a result a tree-like structure is produced illustrating a series of IF/ THEN rules in the following form:

"If *feature1 is good* and … and *featureN is medium* then *case0 belongs to class0*"

THE MEDICAL CASE

Our example derives, as already mentioned, from the medical field. Acute stroke is a common and serious illness causing life-threatening problems to thousands of people every year. The various interpretations of stroke, which are dealt with in this paper, are *large vessel atherosclerosis*, *cardioembolic stroke*, *lacune*, *infarcts of unknown cause*, and *intracerebral hemorrhage*. These five categories will be considered as the classes in our example. The cases that will be assessed comprise of about 170 different attributes defining the patient's condition. In the process of treating stroke cases it is important to have a fast and accurate estimation of the problem. Hence it becomes clear that a manual evaluation can be rather difficult and complicated, which makes obvious the necessity of a decision support system. In

the comparison that will follow we will apply the machine learning methods, which were described in the previous paragraph, in order to support the doctors' decision making and find out the one that best applies to the medical diagnosis of stroke. The automated estimation process will take place in two stages; first we will obtain a primary diagnosis classification and then we will use it to extract the final diagnosis. This shall be repeated for all selection measures in order to acquire a complete picture of the effectiveness of each one.

In our example, the path leading to a leaf takes the following form:

"If *Arrhythmia is Not Apparent* and *Course is Maximal at Onset* and *Age is lower than 77.5 years* and *Glasgow Coma Scale degree is lower than 14.5* and *Systolic Blood Pressure is lower than 155 mmHg* and *History of Cardiac Disease is apparent* and *Stroke took place during Hospitalization* then *Primary Diagnosis is Cardioembolic Stroke*"

At this moment neurology experts are examining the tree structure and the rules for validation purposes in terms of complexity, comprehensibility, correctness and completeness.

SELECTION MEASURES

In this section we will present the selection criteria with a short description for each one and some tests we conducted in order to induct the decision tree structure for each measure. Each criterion is being tested on the training data in order to examine its accuracy degree what the expert's decision methodology is concerned.

ENTROPY BASED MEASURES

Information gain. This selection measure was developed by Quinlan (1986, 1993). It measures the amount of information needed to identify the class of a case belonging to a specified group and then calculates the I_{gain}, which is based on the, by Shannon (1948) defined, entropy *H*. Its major disadvantage is the preference it shows for attributes with many values.

Information gain ratio. This is a normalized version of information gain in order to avoid the bias in favor of attributes with many values. The entropy of the frequency distribution is now shared between the values of the attributes in order to define the gain in useful information.

Symmetric information gain. It was developed by Lopez de Mantaras (1991) and it represents the symmetric version of the previous measure. There are two alternatives: one results by dividing with the common entropy of the classes and the cases and the other by dividing with their sum.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
Information Gain	**79,394**	**77,674**
Information Gain Ratio	**86,13**	**86,812**
Symmetric Information Gain Ratio 1	**84,12**	**85,358**
Symmetric Information Gain Ratio 2	**84,12**	**84,839**

Gini-Index. This measure is based on the so-called quadratic entropy, deriving from the generalized entropy form described by Daroczy (1970). Its use is analogous to the *H* entropy as mentioned above.

Symmetric Gini-Index. This is a symmetric version of Gini-Index by Zhou and Dillon (1991) in order to avoid the preference for attributes with many values.

Modified Gini-Index. It was presented by Kononenko (1994, 1995) and it is an alternative option for reducing the preference for attributes with many values. This measure has a better representation of features with few values.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
Gini Index	**80,1**	**78,193**
Symmetric Gini Index	**84,623**	**84,943**
Modified Gini Index	**80,502**	**77,362**

STATISTICAL MEASURES

Relief measure. Kira and Rendell (1992) as well as Kononenko (1994, 1995) mention this measure. Closely related to Gini-Index, it evaluates attributes according to the correspondence of their values with the classification result. This means that cases with different attribute values have little chance to belong to the same class and the other way around.

$\div^2$ *measure.* Another selection measure deriving from the statistics theory. The chi quadrate used for the evaluation of the various values is termed as the weighted sum of the distance quadrate of deviation.

Evidence Weight. This selection measure compares the odds for a class with and without prior knowledge of the value of an attribute. Then, it judges in favor of the attribute leading to the greater average change of the odd.

Relevance. Baim (1988) implemented this measure which is capable of distinguishing values clearly connected to a class' definition. The bigger the value of this measure the better the connection between the attribute and the classes.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
Relief Measure	**85,427**	**86,916**
$\div^2$ Measure	**80,804**	**77,57**
Evidence Weight	**81,809**	**80,893**
Relevance Measure	**73,768**	**76,428**

BAYES MEASURES

K2 metric. Based on the g function of the Bayesian network theory. The limitation is that all leaves in a specific tree territory are described by the same distribution so all decisions are made in the same way. One more disadvantage is that it's not likelihood equivalent.

Bdeu metric. Is the symmetric form of the previous measure and leads to slightly different evaluation of the attributes.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
K2 Metric	**84,623**	**86,812**
BDeu Metric	**49,145**	**33,229**

MINIMUM DESCRIPTION LENGTH

MDL with coding based on relative frequencies. Information gain takes the coding schema of the information sent for granted. This measure calculates the cost of the coding schema information using relative frequencies.

MDL with coding based on absolute frequencies. Same as the latter using absolute instead of relative frequencies.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
Reduction of Desciption Length 1	**76,582**	**81,828**
Reduction of Desciption Length 2	**78,693**	**85,981**

POSSIBILISTIC MEASURES

Specificity gain. This measure is based on the possibility theory in which nonspecifity plays the same role with the entropy in the probability theory. Gebhardt and Kruse (1996) presented it as a measure for learning possibilistic networks and it is working under the same idea information gain does.

Specificity gain ratio. As by information gain this is a normalized version of the previous measure.

Selection Measure	Primary Diagnosis (% success)	Final Diagnosis (% success)
Specificity Gain	**89,346**	**89,304**
Specificity Gain Ratio	**85,527**	**86,604**
Symmetric Specificity Gain Ratio 1	**85,226**	**87,124**
Symmetric Specificity Gain Ratio 2	**86,13**	**87,643**

CONCLUSIONS

The accuracy we achieved on the data is based on a certain induction schema, which we apply in order to receive an acceptable rule basis. This results in a reduced efficiency level that, on the other hand, offers a less expanded form of the tree and a more reliable set of rules. On the tables presented in the previous paragraph we can easily compare the alternative selection criteria. The most effective among them proved to be the *Specificity Gain*, which resulted in an 89,3% success in recognizing the correct diagnosis. The other specificity measures varied among 86,6 and 87,6% of successful diagnoses. All, except Specificity Gain, showed some improvement during the Final Diagnosis stage. While *Information Gain* poorly appreciated the correct patient's state, its *Ratio* variation (normalization) achieved a percent of 86,8 while the *Symmetric* versions followed with up to two units lower. Three other measures that followed Specificity Gain were the statistical *Relief* measure, the *Reduction of Description Length 2* measure and the *K2 Metric*, which belongs to the Bayes category.

We should here add that the expert's prediction results on new data (we compare the PD and FD parameters) reach a 66 % accuracy success rate. Our classification's success rates in the examination of new data vary between 50% and 75%. We are confident that after further correcting evaluations from our experts these results will show remarkable improvement.

ACKNOWLEDGEMENTS

MIT Gmbh, Aachen, Germany, is greatly acknowledged for providing us with the DataEngine software and plugins, which we used to conduct our experiments.

REFERENCES

Baim, P.W., 1988, "A Method for Attribute Selection in Inductive Learning Systems", IEEE Transcripts on Pattern Analysis and Machine Intelligence, PAMI-10, pp. 888-896.

Darüczy, Z., 1992, "Generalized Information Functions", Information and Control 10, pp. 309-347, Kluwer.

Gebhardt, J.; Kruse, R., 1996, "Tightest Hypertree Decompositions of Multivariate Possibility Distributions", Proc. Int. Conf. On Information Processing and Management of Uncertainty in Knowledge-based Systems.

Kira, K.; Rendell, L., 1992, "A Practical Approach to Feature Selection", Proc. 9th Int. Conf. On Machine Learning, pp. 250-256, Morgan Kaufman, San Franscisco/CA, USA.

Kononenko, I, 1994, "Estimating Attributes: Analysis and Extensions of Relief", Proc. 7th Europ. Conf. On Machine Learning, Springer, New York/NY, USA.

Kononenko, I, 1995, "On Biases in Estimating Multi-Valued Attributes", Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, pp. 1034-1040, Montreal, Canada.

Lopez de Mantaras, R, 1991, "A Distance-based Attribute Selection Measure for Decision Tree Induction", Machine Learning 6, pp. 81-92, Kluwer

Quinlan, J.R., 1986, "Induction of Decision Trees", Machine Learning 1, pp. 81-106, Kluwer

Quinlan, J.R., 1993, "C4.5: Programs for Machine Learning", Morgan Kaufman, San Franscisco/CA, USA.

Shannon, C.E., 1948, "A Mathematical Theory of Communication", The Bell Systems Technical Journal 27, 379-423, 623-656

Zhou, X., Dillon, T.S., 1991, "A Statistical-Heuristic Feature Selection Criterion for Decision Tree Induction", IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-13, 834-841.