Early Machine Learning Research in Ljubljana

Igor Kononenko
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
E-mail: igor.kononenko@fri.uni-lj.si

Keywords: machine learning, decision trees, naive Bayesian classifier, ReliefF

Received: October 17, 2017

We describe early machine learning research in Ljubljana, motivated by medical diagnostic problems, in the areas of building decision trees with Assistant, the development of Naïve and Semi-Naïve Bayesian classifier and its explanations of individual predictions, and the development of ReliefF and RReliefF algorithms for non-myopic evaluation of attributes in classification and regression, respectively.

Povzetek: V članku opišemo zgodnje raziskave na področju strojnega učenja v Ljubljani, ki so bile motivirane z medicinskimi diagnostičnimi problemi. Razvili smo sistem Asistent za gradnjo odločitvenih dreves, naivni in delno naivni Bayesov klasifikator in metodo razlage njunih napovedi ter algoritma ReliefF in RReliefF za nekratkovidno ocenjevanje atributov v klasifikaciji in regresiji.

1 Introduction

As a young researcher, I started my research in Machine learning (ML) in 1982 at the University of Ljubljana and with strong connection with the Artificial Intelligence (AI) group at Jožef Stefan Institute in Ljubljana, Slovenia. My supervisor Prof. Ivan Bratko suggested to me to use Quinlan's (1979) algorithm ID3 for learning medical diagnostic rules. My first data set, obtained from Ljubljana Institute of Oncology, was a description of 339 patients with known correct locations of the primary tumor in the body out of 22 possible locations. The diagnostic task was to determine the location of the primary tumor for new patients, given the description of patients' age, sex, tumor grade, and locations of detected metastases. We tested the classification accuracy of physicians-experts and they were able to correctly classify 42% of patients. The performance of ID3 on this hard diagnostic problem was not satisfactory (lower than 40%), that is why we started to research the possible deficiencies of ID3 and search for the methodologies which would circumvent them.

At that time only few researchers applied ML to medical diagnosis, see (Kononenko, 2001) for an overview. ID3 was developed in 1979 and was not yet applied to medical diagnosis, nobody was using Naïve Bayes (Good, 1950; 1964), which was yet to be rediscovered by us and subsequently by ML community, and more advanced ML approaches, such as multilayered neural networks, support vector machines and random forests were developed much later. Therefore, building decision trees with ID3 seemed to be a good starting point. Note also that there was no internet at that time and the spreading of news about scientific development was significantly slower compared to nowadays. For example, we became aware of system CART (Breiman et al. 1984) for building classification and regression trees several years after it was published.

2 Induction of decision trees with Assistant

Our first discovery was that Information gain, used by ID3 to evaluate the quality of attributes, was biased to overestimate the multivalued attributes, so normalization was required. Another observation was, that lower levels of the tree become unreliable due to small numbers of training examples, so a kind of pruning was needed. Also, at certain level of the tree, built by ID3, a null (empty) leaves could appear, indicating that there was no corresponding training instances for such a leaf, which required a technique to classify new instances which fall in such a leaf. Yet another problem was that ID3 was not able to deal with missing values of attributes. Introduction of an additional value "unknown" for each attribute did not work well, as it led to larger trees and an additional reduction of the number of instances in the leaves.

The research resulted in the development of a new decision tree learning algorithm, called Assistant (Kononenko et al., 1984), which reached the classification accuracy of 44% in the primary tumor diagnostic task.

The reason for encouraging results is that (good) ML algorithms can model the probability distributions more accurately than human experts. On the other hand, physicians use additional information about patients which cannot be straightforwardly coded in a form suitable for ML. Therefore, the comparison of prediction performance is biased, as physicians were, for the sake of comparison, constrained to use the same information as ML algorithms. Our encouraging results motivated other researchers to apply ML in various areas of medical diagnosis, see an overview in (Kononenko, 2001).

The main five contributions of Assistant with respect to ID3 were:

2.1. *An ad-hoc normalization of the Information gain* – dividing information gain of the attribute with k possible values with $\log_2 k$ in order to prevent the overestimation of multivalued attributes. Although it improved the performance, it was ad-hoc. Ross Quinlan, inspired by our research, introduced another normalization – so called Gain-ratio in his famous system C4.5 (Quinlan, 1986), while the appropriate normalization of Information gain was introduced in ML community later with the so-called Distance measure (Mantaras, 1989).

2.2. *Using (an ad-hoc) decision tree pruning*. We introduced a parameter which indicated how many training instances should be in the leaf in order to allow further subtree building. Later, inspired by our idea, many researchers proposed various pre- and post-pruning techniques, however all of them introduced one or more parameters for controlling the strength of pruning. For example, our colleague from Jožef Stefan Institute in Ljubljana, Bojan Cestnik developed a post-pruning technique based on the *m*-estimate of probabilities (Cestnik and Bratko, 1991) which uses parameter *m* for pruning control.

We were looking for a parameter-less pruning techniques, yet without success. We needed another ten years to develop a satisfactory decision tree pre-pruning method which required no parameter setting. The method is based on the MDL-principle (Li and Vitanyi, 1993), which we first used to develop the MDL attribute evaluation method (Kononenko, 1995). The basic idea is to evaluate how compressive a (discrete) attribute is. The effectiveness of that method depends on the appropriate selection of (optimal) data coding. The same idea was later extended to parameter-less decision tree pre-pruning (Kononenko, 1998). The method evaluates how compressive the subtree is in comparison to a leaf alone (without the subtree). Again, the effectiveness of the method depends on the appropriate coding of the data and the tree structure.

2.3. *Classification in combination with the Naïve Bayesian classifier (NB) in the tree leaves.* One version of this idea is to use NB in the empty (null) leaves. This allows us to classify new instances for which no support from the training set in the corresponding leaf exists. The obvious generalization is to use NB in all leaves, allowing the classification process to efficiently use the information of attributes, not tested on the path from the root to the leaf. Later, the same idea was used by researchers who developed regression trees, where in the leaves Linear regression can be used.

2.4. *Building binary decision trees.* In order to avoid over-splitting the training data set (and also to overcome the bias of Information gain to overestimate multivalued attributes) we introduced the binarization of continuous and discrete attributes in order to build binary decision trees. Binary trees proved to be smaller and more accurate, avoiding also the so called replication problem – the appearance of more identical or similar subtrees in a non-binary decision tree.

2.5. *Dealing with incomplete data*. We introduced the methodology for dealing with missing values of attributes, by introducing the instance weights which correspond to the (conditional) probabilities that the instance with missing value has a certain attribute value. The weighted instance then follows all the branches from the current node, each with an appropriate weight. This attribute weighting was generalized to the so called "don't care" values, where any attribute value is allowed. For such an instance the weight is multiplied with the number of possible values of the attribute with "don't care" value. The methodology was later adopted as a standard way for dealing with incomplete data in decision tree learning.

Later, a reimplementation of Assistant was developed, called Assistant 86 (Cestnik et al., 1987) which was followed by a commercial system Assistant Professional.

3 Naïve Bayesian classifier

During the development of the Assistant learning algorithm, I intuitively developed a »simple statistical method«, as I called it at that time and compared its results with decision trees. The surprisingly simple method performed on the primary tumor problem equally well as Assistant did. At that time, however, we claimed that decision trees are preferable due to their "transparency", which does not hold for »statistical methods«. I knew, that my »statistical method« was ad-hoc but I was not able to formally interpret it. With the help of Prof. Bratko we realized that my ad-hoc statistical method was almost the same as the Naïve Bayesian classifier (NB), however lacking the prior probability of the class in the NB formula. (At that time we called it Simple Bayes and only at the ISSEK Workshop in Bled, Slovenia in 1984, where I for the first time presented Assistant for building decision trees, Prof. Donald Michie tossed the name "*Naïve Bayesian classifier*" – and later this name was accepted by ML community).

It turned out that the corrected NB ("statistical method" upgraded with the prior class probability) was able to significantly outperform Assistant in the primary tumor domain (reaching 50% of classification accuracy) as well as on two other medical diagnostic problems (lymphography diagnosis and the breast cancer recurrence prediction).

We became motivated to further research NB in relation to decision trees (Kononenko, 1989a), and we developed the explanation method for NB where for each attribute the amount of information for or against the class is provided in the sum of information contributions during the classification process (Kononenko, 1989b). The explanation is obtained by changing probabilities P in the NB formula into information contributions (using $-\log_2 P$). Surprisingly, this explanation turned out to be more intuitive and more transparent to physicians, who claimed that they also sum up the evidence for or against the diagnosis.

In 1988 I was listening to an inspiring talk by Prof. Igor Grabec in Ljubljana about artificial neural networks and I decided to do more research in this area. We generalized the Hopfield's (1982) discrete model into Bayesian neural networks, where each neuron in the model uses NB (Kononenko, 1989c), and later in my PhD I generalized it into continuous model. Our generalization of NB to Semi-naïve Bayes (Kononenko, 1991) motivated several researchers to try different approaches to avoid the naivety of NB.

At the same time, in his PhD, Bojan Cestnik developed the *m*-estimate of probabilities, which proved to improve the performance of NB (Cestnik, 1990).

4 ReliefF and RReliefF

In 1992 I attended the ICML conference in Aberdeen in Scotland. The audience was highly impressed by the talk of Prof. Larry Rendell, who described the algorithm RELIEF, developed by his PhD student Kira (Kira and Rendell, 1992). RELIEF is a non-myopic attribute evaluator, i.e. it is able to efficiently evaluate the quality of attributes even if there are strong interactions between attributes. This breakthrough in the field of attribute evaluation lead to the development of ReliefF algorithm (Kononenko, 1994) which was later adopted by the ML community as a standard for evaluating the attributes in classification and many improvements and adaptations of RelieF were developed. ReliefF improved RELIEF in three major directions:

1. *Dealing with noisy data.* RELIEF was sensitive to noise in the data. Instead of searching for each instance one nearest hit (nearest instance from the same class) and one nearest miss (nearest instance form the opposite class), ReliefF searches for *k* nearest hits and *k* nearest misses where *k* is a parameter, set by the user (in the same sense as *k*-NN algorithms deal with noise).
2. *Dealing with multiclass problems.* RELIEF was designed for two-class problems only. ReliefF generalizes to more than two classes by searching for *k* nearest misses from each "opposite" class and appropriately weights the contributions of nearest misses with the prior probabilities of corresponding classes.
3. *Dealing with incomplete data.* RELIEF was designed for complete data, without any missing values. While calculating the distances between instances, ReliefF calculates the contributions of attributes with missing values using the conditional probabilities of values given the class. ReliefF is able to evaluate continuous and discrete attributes for classification. Together with my PhD student Marko Robnik-Šikonja, we developed a regressional version of ReliefF, called RReliefF, which enables the evaluation the quality of discrete and continuous attributes in regression (Robnik-Šikonja and Kononenko, 1997). Note that in regression there are no hits and no misses, as instances do not belong to classes, but rather have real values of regression variable. The basic idea of RReliefF is to use the difference of two instances in regression values to model the "probability that two instances do not belong to the same class".

Together with my PhD student Uroš Pompe, we developed also a variant of Relief which enables the (non-myopic) evaluation of literals in Inductive Logic Programming (ILP) (Pompe and Kononenko, 1998). The basic idea is to make a non-symmetrical evaluation measure, biased towards "positive class", as in ILP only positive examples should be covered by good literals (only a theory for the positive class is built) and negative examples should not be covered by good literals.

5 Conclusion

Our development of ML algorithms was highly motivated by medical diagnostic problems. Our applications started in oncology and later spread to other medical areas, such as prognostics of the femoral neck fracture recovery, rheumatology, diagnosis of lower urinary tract disorders, coronary artery disease, sport injuries etc. The overview of our research of ML for medical diagnosis was described in (Kononenko, 2001), which had a great impact on scientific community. Other, earlier references, with the greatest impact on the ML community, include (Kononenko et al., 1984; Cestnik et al., 1987; Kononenko, 1991; 1994).

The unattained goals of our early ML research, *a general method for explaining individual predictions* in a similar way as the NB's explanations, and *a general method for estimating the reliability of individual predictions* of arbitrary prediction models in classification and regression, were achieved by my PhD students: the former goal by Erik Štrumbelj, and the latter goal by the work of Matjaž Kukar, Zoran Bosnić and Darko Pevec (see the overview by Kononenko et al., 2013).

6 References

[1] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and Regression Trees, Wadsforth International Group.

[2] Cestnik, B., Kononenko, I., Bratko, I. ASSISTANT 86: a knowledge-elicitation tool for sophisticated users. In:Bratko, I., Lavrač, N. Progress in machine learning : proc. of European Working Session on Learning EWSL 87. Sigma Press, 1987, p. 31-45.

[3] Cestnik, B. Estimating probabilities. In: Carlucci A. L. (ed.) Proc. ECAI 90. Pitman. 1990, p.147-149.

[4] Cestnik, B., Bratko, I. On estimating probabilities in tree pruning. In: Proc. EWSL-91: European working session on learning, Porto, Portugal, March 6-8, 1991, Springer. p.138-150.

[5] Good I.J., Probability and the Weighing of Evidence. London: Charles Griffin, 1950.

[6] Good I.J., The Estimation of Probabilities - An Essay on Modern Bayesian Methods, Cambridge: The MIT Press, 1964.

[7] Hopfield. J. J. Neural networks and physical systems with emergent collective computational abilities. Nat. Academy of Sc., 79:2554–2558, 1982.

[8] Kira, K. and Rendell, L. A practical approach to feature selection. In D. Sleeman and P. Edwards, eds, Proc. ICML, Aberdeen, UK, 1992, p. 249–256.
[9] Kononenko, I. ID3, Sequential Bayes, Naive Bayes and Bayesian Neural Networks. Proc. of European Working Session on Learning EWSL 1989, Montpellier: France, Dec. 4-6, 1989a, p.91-98.
[10] Kononenko, I. Interpretation of neural networks decisions, IASTED Int. Conf. Expert systems & apps, Zurich, June 26-29 1989b, pp.224-227.
[11] Kononenko, I. Bayesian Neural Networks, Biological Cybernetics Journal 61: 361-370, 1989c.
[12] Kononenko, I. Semi-naive Bayesian classifier, Proc. of European Working Session on Learning EWSL-91, Porto, March 4-6 1991, p.206-219.
[13] Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. In: Proc. ECML-94, Springer, 1994, p. 171-182.
[14] Kononenko, I. On biases in estimating multi-valued attributes. In: Proc. IJCAI-95: Montréal, Canada, August 20-25, 1995. Volume 2, 1995, p. 1034-1040.
[15] Kononenko, I. The minimum description length based decision tree pruning. In Proc. PRICAI '98: Springer, 1998, p. 228-237.
[16] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. intell. med., 2001, 23(1) 89-109.
[17] Kononenko, I., Bratko, I., Roškar, E.: Experiments in automatic learning of medical diagnostic rules, Proc. ISSEK workshop, Bled, august 1984, p. 1-16.
[18] Kononenko, I. Štrumbelj, E., Bosnić, Z., Pevec, D., Kukar, M., Robnik Šikonja, M. Explanation and reliability of individual predictions. Informatica (Lj.), 2013, 37(1) 41-48.
[19] Li, M. and Vitanyi, P. An Introduction to Kolmogorov Complexity and its Applications. Springer Verlag, 1993.
[20] Mantaras. R. L. ID3 revisited: A distance based criterion for attribute selection. Methodologies for Intelligent Systems, Charlotte, U.S.A, 1989.
[21] Pompe, U., Kononenko, I. Efficient induction and effective use of first-order knowledge. Appl. artif. intell., 1998, vol. 12, no. 5, p. 421-453.
[22] Quinlan J.R. Discovering rules by induction from large collections of examples. Expert systems in the Micro Electronic Age, Edinburgh University, 1979.
[23] Quinlan J.R. Induction of Decision Trees. Machine Learning, 1986, 1(1) 81-106.
[24] Robnik Šikonja, M., Kononenko, I. An adaptation of RELIEF for attribute estimation in regression. Proc. ICML'97, Nashville, July 8-12, 1997, p.296-304.