Expert Systems with Applications 36 (2009) 5264–5271

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Accuracy of machine learning models versus "hand crafted" expert systems – A credit scoring case study

Arie Ben-David [a,*], Eibe Frank [b]

[a] Management Information Systems, Department of Technology Management, Holon Institute of Technology, Holon, Israel
[b] Department of Computer Science, University of Waikato, Hamilton, New Zealand

ARTICLE INFO

Keywords:
Machine learning models
Expert systems
Accuracy
Classification
Regression
Hit ratio
Cohen's kappa
Credit scoring

ABSTRACT

Relatively few publications compare machine learning models with expert systems when applied to the same problem domain. Most publications emphasize those cases where the former beat the latter. Is it a realistic picture of the state of the art?

Some other findings are presented here. The accuracy of a real world "mind crafted" credit scoring expert system is compared with dozens of machine learning models. The results show that while some machine learning models can surpass the expert system's accuracy with statistical significance, most models do not. More interestingly, this happened only when the problem was treated as regression. In contrast, no machine learning model showed any statistically significant advantage over the expert system's accuracy when the same problem was treated as classification. Since the true nature of the class data was ordinal, the latter is the more appropriate setting. It is also shown that the answer to the question is highly dependent on the meter that is being used to define accuracy.

1. Introduction

Emerging technologies are often accompanied by some degree of hype. During the 80s there has been a flood of success stories on how expert systems out-performed human experts. Similar reports were published during the 1990s, when machine learning models became increasingly popular. Very few research publications, however, have compared the capabilities of these two approaches: expert systems and machine learning, when applied to similar problems. Furthermore, as expert systems began to include machine learning components, the boundary between these two approaches began to blur, making a direct comparison between expert systems and machine learning quite difficult.

The term expert system is used here to describe a computerized system that encapsulates human knowledge, without resorting to any machine learning or data mining technique. Such systems are typically built using a process nicknamed "knowledge engineering", in which human experts are being interviewed. The end product of this process is a computer program that tries to mimic the way the experts solve the particular problem, usually in some sort of rule-based form. Machine learning models, on the other hand, usually need only data about past decisions. Given these data, they are supposed to generate a model that effectively solves the problem without any further human assistance.

It is not an easy task to compare expert systems and machine learning. Good commercial machine learning and expert systems are often kept secret, due to fear of competition. Unsuccessful implementations rarely find their way to the literature for obvious reasons. Furthermore, most organizations are not willing to commit sufficient resources for developing both types of implementations for solving the same problem. For these reasons, most publications compare either expert systems or machine learning with human performance. Publications that compare expert systems and machine learning, when applied to the same problem domain, are very rare.

Why is it of interest to compare expert systems with machine learning? Besides academic curiosity, an answer to this question is of importance to any commercial firm and government organization that invests or considers investing money in either or both technologies. While there are many pros and cons of each approach, here we only deal with the accuracy aspect through a test case. The other considerations, though very important, are outside the scope of this research.

This research, thus, does not aim at providing the ultimate answer to the question whether machine learning is preferable to expert systems or the other way around. We were fortunate, though, and had a commercial credit scoring expert system handy with some data that enabled us to test dozens of state-of-the-art machine learning models and compare their accuracy with that of the expert system. To the best of our knowledge this is the first time such a comprehensive comparison study is published in the literature.

* Corresponding author. Tel.: +972 3 5026744/6; fax: +972 3 5026650.
E-mail address: hol_abendav@bezeqint.net.il (A. Ben-David).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2008.06.071

2. Related work

Comparing various aspects of machine learning models and expert systems has long been of interest to researchers from various disciplines such as machine learning, artificial intelligence, and decision making – in particular, how accurate machine learning models are when compared with expert systems. However, there have been very few reports that directly compare the two approaches when applied to the same problem. Most publications compare either expert systems or machine learning with human expert performance.

One of the earliest comparisons between machine learning and expert systems can be found in the landmark paper entitled "Knowledge acquisition by encoding expert rules versus computer induction from examples: A case study of involving soybean pathology" by Michalski and Chilausky (1980). They found that in that particular case, a machine learning model was more accurate than the expert system.

In the area of medical diagnosis, Musen (1989) reported an advantage of an expert system over traditional statistical methods, now being in use in machine learning.

Another well-known comparison was published by Creecy, Masand, Smith, and Waltz (1992). It reports a clear accuracy-wise advantage of a version of the *k*-nearest-neighbor algorithm over an expert system that was in use at the US Census Bureau.

A recent publication published by Daren, Warner, White, and Bessant (2005) also reports an advantage of a machine learning model, a feed-forward neural network, over an expert system in terms of accuracy, in the domain of quality control.

Unlike the above reports, which compare the accuracy of a single expert system with one machine learning model, this research performs a comparison with dozens of such models, making it the most comprehensive, methodological, research published in the literature to date. The findings of this research are not all confirmatory with previous reports. On the contrary, as will shortly be seen, unlike the above reports that claim accuracy-wise advantage of machine learning models over expert systems, no statistically significant clear-cut winner was found in the classification experiment to be described shortly.

3. The experiment

The dataset contained 390 examples of consumer loan credit scoring decisions, made by a real-world expert system in one of Israel's leading financial institutions. Both the expert system's decisions and the correct output were available. The input variables were mixed: some numeric, some ordinal, and some were nominal. The output was ordinal, represented by numbers. It could have been symbolic as well though (i.e., "excellent", "very good", "good".).

The Weka machine learning workbench (Witten & Frank, 2005), which has become a standard benchmarking tool in the machine learning community in recent years, was used throughout this experiment. Version 3.4.7 of Weka was used. All the results are based on 10-fold cross-validation. Unlike machine learning models, which are rebuilt from scratch with every fold, the expert system was not rebuilt each time, since it was a unified, complete structure. However, while the expert system remained fixed, the testing data that were fed into it were identical to the data used to validate the accuracy of the machine learning models. At each fold, the expert system was fed with one-tenth of the testing data, and both the predicted and the true values were recorded. All the machine learning algorithms were used with their default parameter settings, as defined in Weka 3.4.7, to reduce the danger of overfitting due to excessive parameter tuning.

Since only one model out of the dozens currently embedded in Weka uses the order within ordinal class values (OrdinalClassClassifier), and due to the fact that ordinal scales share some numeric properties (i.e., the order) as well as some nominal features (i.e., lack of distance), it was decided to conduct two experiments: A. As a classification problem. B. As a regression one. This way, the number of machine learning models that could be used in the experiment increased significantly, as models that can do both regression and classification are relatively rare. Stratified 10-fold cross-validation was used for the nominal-class version of the experiment, while non-stratified 10-fold cross-validation was used in the regression version. The number of possible class values in the classification experiment was 10.

The major findings of the two experiments are given in the following section. A short description of the models which were used, as well as comments about their Weka implementation and their relative accuracy, follows in the Discussion section.

4. Major findings

The results of the classification experiment are presented first in this section, followed by those of regression.

Most classification-related publications in the machine learning literature consider the hit ratio as the major meter for accuracy. Table 1, therefore, shows the classification experiment results sorted by decreasing order of the average hit ratio. The rank of the model, its name in Weka, the percentage of correct classifications (i.e., the hit ratio), half the width of a 95% confidence interval, and the group to which the model belongs in Weka (usually reflecting the type of

Table 1
Classification accuracy sorted by hit ratio

Rank	Model	Hit %	Hit 95% half width CI	Group
1	NaiveBayes	48.72	5.469	Bayes
2	ConjunctiveRule	47.69	3.789	Rules
3	LWL	47.69	3.586	Lazy
4	AdaboostM1	47.69	3.586	Meta
5	MultiboostAB	47.69	3.586	Meta
6	DecisionStump	47.69	3.586	Trees
7	RBFNetwork	47.44	5.060	Functions
8	OneR	46.92	3.674	Rules
9	SimpleLogistic	46.67	5.455	Functions
10	LMT	46.67	5.455	Trees
11	RepTree	46.41	4.000	Trees
12	BayesNet	46.16	6.642	Bayes
13	LogitBoost	45.64	5.174	Meta
14	DecisionTable	45.64	6.459	Rules
15	SMO	45.38	5.119	Functions
16	Bagging	45.13	4.830	Meta
17	ClassificationViaRegression	44.62	2.476	Meta
18	RandomForest	44.10	4.876	Trees
19	AttributeSelectedClassifier	44.10	3.648	Meta
20	OrdinalClassClassifier	43.85	4.926	Meta
21	EXPERT SYSTEM	43.33	5.763	–
22	Decorate	43.08	3.648	Meta
23	MulltilayerPerceptron	42.31	5.752	Functions
24	NBTree	42.05	5.941	Trees
25	NNGE	41.54	6.574	Rules
26	RandomCommittee	41.28	4.849	Meta
27	FilteredClassifier	40.51	4.952	Meta
28	MulticlassClassifier	40.26	5.334	Meta
29	Logistic	39.74	5.134	Functions
30	J48	39.49	4.752	Trees
31	Ridor	38.97	5.723	Rules
32	Part	37.44	2.623	Rules
33	Kstar	36.67	3.774	Lazy
34	IBK	35.38	4.306	Lazy
35	JRIP	34.36	3.789	Rules
36	RandomTree	33.85	5.101	Trees
37	ZeroR	29.23	0.947	Rules