

# ***EkP*: A Fast Minimization-Based Prototype Selection Algorithm**

Karol Grudziński<sup>1,2</sup>

<sup>1</sup> Department of Physics, Kazimierz Wielki University, Bydgoszcz, Poland

<sup>2</sup> Institute of Applied Informatics, University of Economy, Bydgoszcz, Poland

## **Abstract**

In this paper a completely new system for instance selection which is called SBL-PM-M-*EkP* (*EkP*) is introduced. The study of suitability of the *EkP* method for training data reduction has been studied on 12 datasets. As an underlying classifier the well known IB1 system (1-Nearest Neighbor classifier) has been chosen. We compare the generalization ability of our method to the performance of IB1 trained on the entire training data and the performance of LVQ for which usually the same number of codebooks have been chosen as the number of prototypes which have been selected by the *EkP* system. The results indicate that even with only a few prototypes which have been chosen by the *EkP* method, on all 12 datasets statistically indistinguishable results from those attained with IB1 have been obtained. In several cases the generalization ability of the *EkP* system has been larger than the one attained with LVQ.

## **1 Introduction**

Data mining is commonly used in many domains. A case-based way of data explanation is very popular among researchers. Such an approach to knowledge discovery and understanding is particularly often employed in medicine, where a medical doctor makes a diagnosis by referring to other similar cases in a database of patients.

Interesting instance vectors can be either selected from training data or can be generated out of a training set. In this case the instances' features have in general different values than the ones that are stored in the original training set. Both techniques (i.e. instance selection and prototype generation) usually lead to, often significant, training set size reduction.

This paper concerns the first mentioned above problem, i.e. 'instance selection', 'training data compression, reduction or pruning'. The idea behind this machine learning paradigm is that only a small fraction of a usually much larger, original training set is used for a final classification of unseen samples (Maloof M., Michalski, R. S., 2000; Martinez T. R., Wilson D. R., 1997, 2000; Grochowski M., 2003; Grochowski M. Jankowski N., 2004-1; Grochowski M., Jankowski N., 2004-2; Duch. W., Grudziński. K., 2000; Grudziński K., 2004). A very interesting prototype-rule approach to better data understanding has been made by Duch

et.al. as the alternative to classic rule-based way of data explanation (Duch W., Grudziński K., 2001; Blachnik M., Duch W., 2004).

Prototype selection is an extremely important problem which has been frequently studied by machine learning and pattern recognition researchers. Selection of prototypes can significantly speed up classification and analysis of data later and usually leads to better data understanding and may lower sensitivity to noise of some classifiers. Strong training set reduction may result in, sometimes statistically significant, degradation of the classification accuracy attained on unseen samples, however as many experiments illustrate, often it is the other way around, i.e. data pruning improves generalization ability of classifiers.

The acronym SBL-PM-M-EkP is short for **S**imilarity-**B**ased-**L**earner-**P**artial-**M**emory-**M**inimization-**E**xactly-**k**-**P**rototypes. We want to stress here that our new system differs from our earlier model, SBL-PM-M (Grudziński K., 2004), despite of the similar acronym both systems have been given. In the text we will be omitting the part SBL-PM-M when referring to the SBL-PM-M-EkP method mainly because this algorithm is independent on the classifiers employed and not necessarily has any connection with similarity based methods. SBL is in a way the unofficial trademark of Włodzisław Duch and Karol Grudziński.

## 2 The SBL-PM-M-EkP System

The EkP system is based on minimization of the cost function which returns the number of errors the classifier makes. Despite of this, the EkP method is extremely fast as during every evaluation of the cost function only the preset  $k$  instances are classified. It takes seconds for the EkP method to perform 10-fold cross-validation on most common UCI datasets. In our Weka implementation we used the well known simplex method (Nelder J. A., Mead R., 1965) for a function minimization which we have taken from the Internet (Lampton M., 2004).

The simplex must be initialized first before the minimization procedure is started. The cost function algorithm is given below.

Our implementation of the EkP method is not the simplest one as our code will become a basis for an extended version of this algorithm. In order to give a short description of the algorithm in the text of the paper, it is worth mentioning that the array of optimization parameters is ( $\text{numProtoPerClass} * \text{numClasses} * \text{numAttributes}$ ) dimensional but the instances stored in this vector are not involved in any parameter modification. They are simply being extracted from the parameter vector and are added to the cross-validation training partition in every cost function evaluation. In other words the cross-validation training partitions are built by extracting samples from a parameter vector which always contains  $\text{numProtoPerClass}$  samples from every class which occurs in a problem domain. In a simpler implementation one could store the indexes of the training set instances instead of storing the  $\text{numProtoPerClass} * \text{numClasses}$  vectors themselves in the parameter array.

---

**Algorithm 1** The EkP cost function algorithm
 

---

**Require:** A training set **trainInstances****Require:** A vector  $p[]$  of optimization parameters (**numProtoPerClass** \* **numClasses** \* **numAttributes** dimensional)**for**  $k = 1$  to **numCrossValidationLearningFolds** **do**  Create the empty training set **cvTrain**  Build the  $k$ -th test partition **cvTest**  **for**  $i = 1$  to **numClasses** \* **numProtoPerClass** **do**    **for**  $j = 1$  to **numAttributes** **do**      Add the prototype stored in  $p[]$  starting from  $p[j + \text{numAttributes} * i]$   
      and ending in  $p[j + \text{numAttributes} + \text{numAttributes} * i]$  to **cvTrain**    **end for**  **end for**  Build (train) the classifier on **cvTrain** and test it on **cvTest**  **end for**Remember the optimal  $p[]$  value and the associated with it lowest value of **numClassificationErrors****return** **numClassificationErrors**


---

### 3 Numerical Experiments

In order to verify suitability of the EkP system for data analysis the classification experiments on 12 real-world problems (mainly taken from the well-known UCI repository of machine-learning databases (Mertz C. J., Murphy P. M.)) have been performed. The information about the datasets used can be found in Table 1.

TABLE 1: Datasets used in our experiments

| #   | Name                    | # Instances | # Attributes | # Numeric | # Nominal | Class Information                                | Missing Val.                                    |
|-----|-------------------------|-------------|--------------|-----------|-----------|--|---|
| 1.  | appendicitis            | 106         | 7            | 7         | 0         | 1. 85 (80.2%)<br>2. 21 (19.8%)                   | 1<br>2<br>none                                  |
| 2.  | balance-scale           | 625         | 4            | 4         | 0         | 1. 49 (7.8%)<br>2. 288 (46.1%)<br>3. 288 (46.1%) | Balanced<br>Left<br>Right<br>none               |
| 3.  | breast-cancer           | 286         | 9            | 9         | 0         | 1. 201 (70.3%)<br>2. 85 (29.7%)                  | no-recurrence events<br>recurrence-events<br>9  |
| 4.  | breast-cancer-Wisconsin | 699         | 9            | 9         | 0         | 1. 458 (65.5%)<br>2. 241 (34.5%)                 | no-recurrence events<br>recurrence-events<br>16 |
| 5.  | german credit           | 1000        | 20           | 7         | 13        | 1. 700 (70.0%)<br>2. 300 (30.0%)                 | good<br>bad<br>none                             |
| 6.  | credit rating           | 690         | 15           | 6         | 9         | 1. 307 (44.5%)<br>2. 383 (55.5%)                 | +<br>-<br>37                                    |
| 7.  | Cleveland heart         | 303         | 13           | 6         | 7         | 1. 165 (54.5%)<br>2. 138 (45.5%)                 | < 50<br>> 51<br>7                               |
| 8.  | heart Hungarian         | 294         | 13           | 6         | 7         | 1. 188 (63.9%)<br>2. 106 (36.1%)                 | < 50<br>> 51<br>782                             |
| 9.  | heart statlog           | 270         | 13           | 13        | 0         | 1. 150 (55.6%)<br>2. 120 (44.4%)                 | absent<br>present<br>none                       |
| 10. | hepatitis               | 155         | 19           | 6         | 13        | 1. 32 (20.6%)<br>2. 123 (79.4%)                  | die<br>live<br>167                              |
| 11. | iris                    | 150         | 4            | 4         | 0         | 1. 50 (33.3%)<br>2. 50 (33.3%)<br>3. 50 (33.3%)  | setosa<br>versicolor<br>virginica<br>none       |
| 12. | pima-diabetes           | 768         | 8            | 8         | 0         | 1. 500 (65.1%)<br>2. 268 (34.9%)                 | negative<br>positive<br>none                    |

#### 3.1 The Description of the Conducted Experiments

The EkP system can be based on an arbitrary classifier, i.e. it can be a neural-network, support-vector machine or a decision-tree method, etc. In our experiments the IB1 (Aha D. W., Albert M. K., Kibler D. , 1991) system has been

used both as the underlying classifier for the *EkP* system and as the reference method. The reason for selecting the IB1 system is that this method requires very small training datasets which may consist of just a few samples in order to make classification possible. Other classifiers, including *IBk* (Aha D. W, Albert M. K., Kibler D. , 1991) require slightly larger training sets in order to operate. Our aim when we were conducting the experiments for this paper was to show that even the calculations with the extremely low number of prototypes selected may lead to attaining excellent results on unseen samples. The well known LVQ (Hyninen, Kangas, Kohonen, Laaksonen, Torkolla, 1996; Kohonen T., 2001; Kaski S., Kohonen T. Oja M., 2003) method which is however a prototype-generation system has also been taken as the reference model in our experiments. The second reason for choosing the IB1 classifier as the underlying method for the *EkP* system is the fact that the LVQ method uses the *k*-Nearest Neighbor classifier as its classification engine.

Generalization ability of the *EkP* system with only 1, 2 and 3 instances per class selected from the training set has been compared to the classification performance of LVQ for which usually the same number of codebooks has been used.<sup>1</sup> Additionally, the results obtained with the IB1 (1-NN) system which was trained on the entire training partitions and those attained with the majority classifier (ZeroR) are provided.

The 10-fold stratified cross-validation test has been performed for all 12 domains. In the experiments conducted with LVQ and *EkP* systems, in each cross-validation fold, the training partition has been pruned so that only the prototype cases remained and the generalization ability has been estimated on the cross-validation test partition. After the completion of the calculation on all 10 folds the test has been repeated 10 times and the average classification accuracy and its standard deviation which were taken over the all available 100 partial results have been reported.

The corrected re-sampled T-Test (Frank E., Witten I.H., 2000; Dobosz K. , 2006) has been used to calculate statistical significance of the results (with the factor of 0.05) in order to help making the decision whether the *EkP* system performed better, the same or worse than the reference models.

In the all experiments LVQ, version 1, with 2, 4, 6, codebooks, 'Random Training Data Proportional' initialization, learning rate of 0.3, total training iterations of 1000 and disabled voting has been used. These are the default parameters for this classifier. The LVQWeka implementation of the LVQ method that has been employed in our calculations was written by Jason Brownlee (Brownlee J., 2004).

The *EkP* system has also been used mainly with the default settings and trained with the simplex minimization restricted to 300 cost evaluations. This value was

---

<sup>1</sup>In case of three class problems the number of prototypes retained by *EkP* is slightly larger than those generated by the LVQ system. The reason for this is that the calculations have been performed in a batch mode with the same settings for all 12 domains. This made collecting the results and typesetting the paper much easier, particularly the tables. *EkP* takes the number of instances per class to be retained as its adaptive parameter whilst in the case of LVQ, the total number of codebooks to be used has to be specified. Thus in the experiments on two class problems with both LVQ and *EkP*, 2, 4 and 6 prototypes have been retained while for three class problems the LVQ system was used with 2, 4, and 6 codebooks but our *EkP* method with 3, 6 and 9 prototypes.

taken as the default for the classifiers of the size of a couple of hundred cases and this choice is based on our earlier experience with similar minimization-based learning systems we had been working on. Two experiments have been conducted with 60 and 10 simplex points respectively. The latter value is the default. It seems that the larger is the number of simplex points, the better results are obtained. The upper limit on the number of simplex points is the number of samples in the training partition. Leave-one-out learning and the IB1 classifier which has been chosen as the classification engine are the parameters with which our method has been used on all 12 classification problems. The influence of the selection of the value of cross-validation learning fold has not been yet investigated. We know only that leave-one-out learning seems to provide the best way to obtain good generalization at the expense of significantly lengthening the calculation time. Finally, what remains to be mentioned is, that the *EkP* system has been written by the author in the Java programming language as a plugin to the well known Weka machine learning workbench (Frank E., Witten I.H., 2000).

### 3.2 Discussion of the Results of the Experiments

Two experiments with sixty and ten simplex points initialized by the *EkP* system have been performed in order to make a first step in estimating the influence of the number of them on the generalization ability and the time requirements of our method.

What is common to both experiments is that eight results out of twelve obtained with the IB1 system have shown statistically significant improvement over the majority classifier. LVQ for which two codebooks have been used performed very poorly and outperformed the majority classifier only once. LVQ for which four and six codebooks were chosen attained statistically better results in classification than the majority classifier five and six times respectively. The LVQ method with two codebooks used seven times obtained statistically worse results than those attained by IB1, however with four and six codebooks – only five times.

TABLE 2: 10-Fold cross-validation results obtained on the selected datasets with the IB1 system trained on the whole training partitions, LVQ with 2, 4, 6 codebooks and SBL-PM-M-*EkP* with 1, 2, 3 prototypes per class vs. the ZeroR (i.e. majority) classifier. 60 simplex points have been chosen for the *EkP* system.

| Data Set                | ZeroR      | IB1         | LVQ-2       | LVQ-4       | LVQ-6       | EkP-1       | EkP-2      | EkP-3      |
|-------------------------|------------|-------------|-------------|-------------|-------------|-------------|------------|------------|
| appendicitis            | 80.18±2.64 | 80.28±10.78 | 78.64±15.17 | 82.72±10.92 | 85.15±9.37  | 86.25±10.16 | 86.37±9.20 | 87.05±8.95 |
| balance-scale           | 45.76±0.53 | 78.16±4.98  | 63.76±21.19 | 80.81±14.37 | 84.51±8.40  | 79.51±4.79  | 79.73±5.01 | 80.46±4.64 |
| breast-cancer           | 70.30±1.37 | 68.58±7.52  | 66.46±12.18 | 70.97±4.10  | 71.00±4.79  | 73.55±6.39  | 73.01±5.76 | 73.09±5.21 |
| wisconsin-breast-cancer | 65.52±0.44 | 95.45±2.52  | 75.01±23.75 | 90.10±13.65 | 93.12±9.00  | 95.98±2.26  | 96.10±2.30 | 96.20±2.30 |
| german-credit           | 70.00±0.00 | 71.88±3.68  | 66.84±10.94 | 69.57±4.20  | 69.78±1.54  | 69.70±1.49  | 69.83±3.12 | 69.44±3.22 |
| credit-rating           | 55.51±0.67 | 81.57±4.57  | 53.04±5.39  | 58.72±5.18  | 62.35±5.08  | 79.62±6.13  | 81.41±5.02 | 82.51±4.55 |
| Cleveland-heart         | 54.45±1.29 | 76.06±6.84  | 56.52±8.48  | 62.14±8.94  | 62.77±8.01  | 80.34±7.17  | 80.46±6.68 | 79.55±7.11 |
| Hungarian-heart         | 63.95±1.36 | 78.33±7.54  | 62.15±13.00 | 67.42±9.39  | 65.88±6.80  | 82.45±6.51  | 81.98±7.08 | 82.26±6.73 |
| heart-statlog           | 55.56±0.00 | 76.15±8.46  | 56.89±8.23  | 62.30±8.40  | 64.41±8.55  | 81.04±6.71  | 80.00±7.33 | 80.11±7.48 |
| hepatitis               | 79.38±2.26 | 81.40±8.55  | 75.39±14.66 | 78.84±3.88  | 77.94±4.99  | 81.74±9.70  | 83.20±9.37 | 83.25±8.88 |
| iris                    | 33.33±0.00 | 95.40±4.80  | 42.27±14.78 | 64.53±25.79 | 78.00±23.68 | 91.27±7.68  | 94.53±5.72 | 94.13±5.94 |
| pima-diabetes           | 65.11±0.34 | 70.62±4.67  | 61.96±9.75  | 66.79±4.32  | 68.46±5.22  | 70.41±5.90  | 71.08±5.67 | 71.13±4.89 |
| average                 | 61.59±0.91 | 79.49±6.24  | 63.24±13.13 | 71.24±9.43  | 73.61±7.95  | 80.97±6.24  | 81.48±6.02 | 81.60±5.83 |

◦, • statistically significant improvement or degradation

What concerns the *EkP* system, in the first experiment (60 simplex points), statistically significant improvement over the majority classifier has been attained by our method with one and two prototypes per class selected exactly eight times and on the same datasets as in the case of IB1. *EkP* with three prototypes per class selected outperformed the majority classifier nine times. All experiments with the

$EkP$  method indicated statistical insignificance of the results of the generalization with respect to IB1.

TABLE 3: 10-Fold cross-validation results obtained on the selected datasets with the ZeroR system (i.e. majority classifier), LVQ with 2, 4, 6 codebooks and SBL-PM-M- $EkP$  with 1, 2, 3 prototypes per class vs. the IB1 classifier trained on the whole training partitions. 60 simplex points have been chosen for the  $EkP$  system.

| Data Set                | IB1         | ZeroR       | LVQ-2        | LVQ-4        | LVQ-6        | $EkP-1$     | $EkP-2$     | $EkP-3$    |
|-------------------------|-------------|-------------|--------------|--------------|--------------|-------------|-------------|------------|
| appendicitis            | 80.28±10.78 | 80.18±2.64  | 78.64±15.17  | 82.72±10.92  | 85.15±9.37   | 86.25±10.16 | 86.37±9.20  | 87.05±8.95 |
| balance-scale           | 78.16±4.98  | 45.76±0.53● | 63.76±21.19  | 80.81±14.37  | 84.51±8.40   | 79.31±4.79  | 79.73±5.01  | 80.46±4.64 |
| breast-cancer           | 68.58±7.52  | 70.30±1.37  | 66.46±12.18  | 70.97±4.10   | 71.00±4.79   | 73.55±6.39  | 73.01±5.76  | 73.09±5.21 |
| wisconsin-breast-cancer | 95.45±2.52  | 65.52±0.44● | 75.01±23.75● | 90.10±13.65  | 93.12±9.00   | 95.98±2.26  | 96.10±2.30  | 96.20±2.30 |
| german-credit           | 71.88±3.68  | 70.00±0.00  | 66.84±10.94  | 69.57±4.20   | 69.78±1.54   | 69.70±1.49  | 69.83±3.12  | 69.44±3.22 |
| credit-rating           | 81.57±4.57  | 55.51±0.67● | 53.04±5.39●  | 58.72±5.18●  | 62.35±5.08●  | 79.62±6.13  | 81.41±5.02  | 82.51±4.55 |
| Cleveland-heart         | 76.06±6.84  | 54.45±1.29● | 56.52±8.48●  | 62.14±8.94●  | 62.77±8.01●  | 80.34±7.17  | 80.46±6.68○ | 79.55±7.11 |
| Hungarian-heart         | 78.33±7.54  | 63.95±1.36● | 62.15±13.00● | 67.42±9.39●  | 65.88±6.80●  | 82.45±6.51  | 81.98±7.08  | 82.26±6.73 |
| heart-statlog           | 76.15±8.46  | 55.56±0.00● | 56.89±8.23●  | 62.30±8.40●  | 64.41±8.55●  | 81.04±6.71  | 80.00±7.33  | 80.11±7.48 |
| hepatitis               | 81.40±8.55  | 79.38±2.26  | 75.39±14.66  | 78.84±3.88   | 77.94±4.99   | 81.74±9.70  | 83.20±9.37  | 83.25±8.88 |
| iris                    | 95.40±4.80  | 33.33±0.00● | 42.27±14.78● | 64.53±25.79● | 78.00±23.68● | 91.27±7.68  | 94.53±5.72  | 94.13±5.94 |
| pima-diabetes           | 70.62±4.67  | 65.11±0.34● | 61.96±9.75●  | 66.79±4.32   | 68.46±5.22   | 70.41±5.90  | 71.08±5.67  | 71.13±4.89 |
| average                 | 79.49±6.24  | 61.59±0.91  | 63.24±13.13  | 71.24±9.43   | 73.61±7.95   | 80.97±6.24  | 81.48±6.02  | 81.60±5.83 |

○, ● statistically significant improvement or degradation

The training times of the  $EkP$  system, which are however all statistically worse than those of IB1 (it is not a surprise), are quite short and in average are equal to about 4s for learning on a single partition of a dataset of the size of a couple of hundred cases. This makes less than a minute for the entire 10-fold cross-validation test to complete on most common UCI datasets. It should be noted that training the  $EkP$  method with lower-fold cross-validation than leave-one-out leads to a significant reduction of the time requirements for this algorithm.

TABLE 4: Average time in seconds of training the classifiers used in the experiment 1 on a single cross-validation fold. Statistical significance with the factor of 0.05 has been calculated with respect to the training times of IB1. 60 simplex points have been chosen for the  $EkP$  system.

| Data Set                | IB1       | ZeroR     | LVQ-2     | LVQ-4     | LVQ-6     | $EkP-1$     | $EkP-2$     | $EkP-3$     |
|-------------------------|-----------|-----------|-----------|-----------|-----------|-------------|-------------|-------------|
| appendicitis            | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.00 | 0.00±0.01 | 0.31±0.10●  | 0.37±0.15●  | 0.45±0.11●  |
| balance-scale           | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 4.49±0.24●  | 5.00±0.88●  | 5.21±0.53●  |
| breast-cancer           | 0.00±0.00 | 0.00±0.00 | 0.00±0.03 | 0.00±0.03 | 0.00±0.01 | 1.23±0.22●  | 1.48±0.28●  | 1.71±0.33●  |
| wisconsin-breast-cancer | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 5.77±0.31●  | 6.66±0.84●  | 7.19±0.31●  |
| german-credit           | 0.00±0.00 | 0.00±0.00 | 0.01±0.01 | 0.01±0.01 | 0.01±0.01 | 13.15±1.08● | 15.58±0.30● | 18.16±1.09● |
| credit-rating           | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.01±0.03 | 6.27±0.34●  | 7.49±0.34●  | 8.62±0.30●  |
| Cleveland-heart         | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 2.33±0.31●  | 3.48±0.27●  | 4.62±0.30●  |
| Hungarian-heart         | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 2.22±0.24●  | 3.32±0.27●  | 4.42±0.29●  |
| heart-statlog           | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 1.40±0.18●  | 1.85±0.34●  | 2.23±0.32●  |
| hepatitis               | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.01±0.01 | 0.00±0.01 | 0.79±0.17●  | 1.27±0.69●  | 1.56±0.21●  |
| iris                    | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.44±0.14●  | 0.61±0.28●  | 0.62±0.18●  |
| pima-diabetes           | 0.00±0.00 | 0.00±0.02 | 0.00±0.02 | 0.00±0.01 | 0.00±0.01 | 6.76±0.31●  | 7.25±0.37●  | 7.92±0.86●  |
| average                 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 3.76±0.30   | 4.53±0.42   | 5.23±0.40   |

○, ● statistically significant improvement or degradation

In the second experiment substantially lower amount of simplex points (10) has been selected for the calculations that have been conducted with the  $EkP$  method. Statistically significant improvement over the majority classifier has been obtained by the  $EkP$  system with one prototype per class selected seven times which is one time less than in the first experiment.  $EkP$  with two and three prototypes per class chosen also performed excellent and with only 10 simplex points chosen outperformed the majority classifier on eight datasets.

What concerns the statistical significance of the results with respect to IB1, the  $EkP$  method with only one prototype per class performed worse than IB1 only once. In case of two and three samples per class chosen, the results have been statistically insignificant.

Selection of ten instead of sixty simplex points had a strong influence on the time requirements of our method. In this experiment it took only approximately about 10 seconds of CPU time<sup>2</sup> to complete the 10-fold cross-validation test with-

<sup>2</sup>The calculations have been performed under 32-bit Windows XP SP2 operating system,

TABLE 5: 10-Fold cross-validation results obtained on the selected datasets with the IB1 system trained on the whole training partitions, LVQ with 2, 4, 6 codebooks and SBL-PM-M-EkP with 1, 2, 3 prototypes per class vs. the ZeroR (i.e. majority) classifier. 10 simplex points have been chosen for the EkP system.

| Data Set                | ZeroR      | IB1         | LVQ-2       | LVQ-4       | LVQ-6       | EKP-1      | EKP-2      | EKP-3      |
|-------------------------|------------|-------------|-------------|-------------|-------------|------------|------------|------------|
| appendicitis            | 80.18±2.64 | 80.28±10.78 | 78.64±15.17 | 82.72±10.92 | 85.15±9.37  | 85.28±9.57 | 86.61±9.73 | 85.85±9.77 |
| balance-scale           | 45.76±0.53 | 78.16±4.98  | 63.76±21.19 | 80.81±14.37 | 84.51±8.40  | 74.70±5.53 | 76.42±5.40 | 77.72±4.73 |
| breast-cancer           | 70.30±1.37 | 68.58±7.52  | 66.46±12.18 | 70.97±4.10  | 71.00±4.79  | 70.81±7.33 | 72.33±6.46 | 72.80±4.99 |
| wisconsin-breast-cancer | 65.52±0.44 | 95.45±2.52  | 75.01±23.75 | 90.10±13.65 | 93.12±9.00  | 95.37±2.77 | 95.90±2.15 | 95.82±2.26 |
| german-credit           | 70.00±0.00 | 71.88±3.68  | 66.84±10.94 | 69.57±4.20  | 69.78±1.54  | 69.86±0.86 | 69.66±1.88 | 69.77±2.58 |
| credit-rating           | 55.51±0.67 | 81.57±4.57  | 53.04±5.39  | 58.72±5.18  | 62.35±5.08  | 78.19±6.36 | 78.94±5.59 | 80.03±6.24 |
| Cleveland-heart         | 54.45±1.29 | 76.06±6.84  | 56.52±8.48  | 62.14±8.94  | 62.77±8.01  | 80.01±6.88 | 79.87±7.23 | 79.24±6.89 |
| Hungarian-heart         | 63.95±1.36 | 78.33±7.54  | 62.15±13.00 | 67.42±9.39  | 65.88±6.80  | 83.11±6.57 | 82.25±6.27 | 81.81±6.61 |
| heart-statlog           | 55.56±0.00 | 76.15±8.46  | 56.89±8.23  | 62.30±8.40  | 64.41±8.55  | 78.93±7.49 | 78.63±6.63 | 78.48±7.75 |
| hepatitis               | 79.38±2.26 | 81.40±8.55  | 75.39±14.66 | 78.84±3.88  | 77.94±4.99  | 79.28±8.06 | 80.87±8.65 | 81.12±8.57 |
| iris                    | 33.33±0.00 | 95.40±4.80  | 42.27±14.78 | 64.53±25.79 | 78.00±23.68 | 88.27±7.99 | 91.60±6.81 | 93.93±6.50 |
| pima-diabetes           | 65.11±0.34 | 70.62±4.67  | 61.96±9.75  | 66.79±4.32  | 68.46±5.22  | 68.25±5.83 | 69.50±5.25 | 69.53±5.27 |
| average                 | 61.59±0.91 | 79.49±6.24  | 63.24±13.13 | 71.24±9.43  | 73.61±7.95  | 79.34±6.27 | 80.22±6.00 | 80.51±6.01 |

◦, • statistically significant improvement or degradation

TABLE 6: 10-Fold cross-validation results obtained on the selected datasets with the ZeroR system (i.e. majority classifier), LVQ with 2, 4, 6 codebooks and SBL-PM-M-EkP with 1, 2, 3 prototypes per class vs. the IB1 classifier trained on the whole training partitions. 10 simplex points have been chosen for the EkP system.

| Data Set                  | IB1         | ZeroR      | LVQ-2       | LVQ-4       | LVQ-6       | EKP-1      | EKP-2      | EKP-3      |
|---------------------------|-------------|------------|-------------|-------------|-------------|------------|------------|------------|
| appendicitis              | 80.28±10.78 | 80.18±2.64 | 78.64±15.17 | 82.72±10.92 | 85.15±9.37  | 85.28±9.57 | 86.61±9.73 | 85.85±9.77 |
| balance-scale             | 78.16±4.98  | 45.76±0.53 | 63.76±21.19 | 80.81±14.37 | 84.51±8.40  | 74.70±5.53 | 76.42±5.40 | 77.72±4.73 |
| breast-cancer             | 68.58±7.52  | 70.30±1.37 | 66.46±12.18 | 70.97±4.10  | 71.00±4.79  | 70.81±7.33 | 72.33±6.46 | 72.80±4.99 |
| wisconsin-breast-cancer   | 95.45±2.52  | 65.52±0.44 | 75.01±23.75 | 90.10±13.65 | 93.12±9.00  | 95.37±2.77 | 95.90±2.15 | 95.82±2.26 |
| german-credit             | 71.88±3.68  | 70.00±0.00 | 66.84±10.94 | 69.57±4.20  | 69.78±1.54  | 69.86±0.86 | 69.66±1.88 | 69.77±2.58 |
| credit-rating             | 81.57±4.57  | 55.51±0.67 | 53.04±5.39  | 58.72±5.18  | 62.35±5.08  | 78.19±6.36 | 78.94±5.59 | 80.03±6.24 |
| Cleveland-14-heart-diseas | 76.06±6.84  | 54.45±1.29 | 56.52±8.48  | 62.14±8.94  | 62.77±8.01  | 80.01±6.88 | 79.87±7.23 | 79.24±6.89 |
| Hungarian-14-heart-diseas | 78.33±7.54  | 63.95±1.36 | 62.15±13.00 | 67.42±9.39  | 65.88±6.80  | 83.11±6.57 | 82.25±6.27 | 81.81±6.61 |
| heart-statlog             | 76.15±8.46  | 55.56±0.00 | 56.89±8.23  | 62.30±8.40  | 64.41±8.55  | 78.93±7.49 | 78.63±6.63 | 78.48±7.75 |
| hepatitis                 | 81.40±8.55  | 79.38±2.26 | 75.39±14.66 | 78.84±3.88  | 77.94±4.99  | 79.28±8.06 | 80.87±8.65 | 81.12±8.57 |
| iris                      | 95.40±4.80  | 33.33±0.00 | 42.27±14.78 | 64.53±25.79 | 78.00±23.68 | 88.27±7.99 | 91.60±6.81 | 93.93±6.50 |
| pima-diabetes             | 70.62±4.67  | 65.11±0.34 | 61.96±9.75  | 66.79±4.32  | 68.46±5.22  | 68.25±5.83 | 69.50±5.25 | 69.53±5.27 |
| average                   | 79.49±6.24  | 61.59±0.91 | 63.24±13.13 | 71.24±9.43  | 73.61±7.95  | 79.34±6.27 | 80.22±6.00 | 80.51±6.01 |

◦, • statistically significant improvement or degradation

out repetition, i.e. about five times shorter than in the case of the first experiment.

TABLE 7: Average time in seconds of training the classifiers used in the experiment 2 on a single cross-validation fold. Statistical significance with the factor of 0.05 has been calculated with respect to the training times of IB1. 10 simplex points have been chosen for the EkP system.

| Data Set                  | IB1       | ZeroR     | LVQ-2     | LVQ-4     | LVQ-6     | EKP-1     | EKP-2     | EKP-3     |
|---------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| appendicitis              | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.00 | 0.06±0.06 | 0.07±0.05 | 0.08±0.05 |
| balance-scale             | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.00 | 0.69±0.21 | 0.76±0.21 | 0.82±0.13 |
| breast-cancer             | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.20±0.11 | 0.24±0.12 | 0.28±0.12 |
| wisconsin-breast-cancer   | 0.00±0.03 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 0.90±0.24 | 1.02±0.27 | 1.13±0.21 |
| german-credit             | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.01±0.01 | 0.01±0.01 | 2.01±0.29 | 2.42±0.25 | 3.01±0.77 |
| credit-rating             | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.01±0.01 | 0.01±0.01 | 0.97±0.21 | 1.18±0.23 | 1.36±0.19 |
| Cleveland-14-heart-diseas | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.01±0.03 | 0.00±0.01 | 0.38±0.15 | 0.56±0.16 | 0.89±0.66 |
| Hungarian-14-heart-diseas | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 0.35±0.14 | 0.53±0.17 | 0.70±0.15 |
| heart-statlog             | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 0.23±0.11 | 0.30±0.07 | 0.36±0.11 |
| hepatitis                 | 0.00±0.00 | 0.00±0.00 | 0.01±0.03 | 0.00±0.01 | 0.00±0.01 | 0.13±0.08 | 0.19±0.07 | 0.25±0.09 |
| iris                      | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.08±0.06 | 0.09±0.05 | 0.10±0.05 |
| pima-diabetes             | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 1.03±0.22 | 1.14±0.18 | 1.25±0.21 |
| average                   | 0.00±0.00 | 0.00±0.00 | 0.00±0.01 | 0.00±0.01 | 0.00±0.01 | 0.59±0.16 | 0.71±0.16 | 0.83±0.23 |

◦, • statistically significant improvement or degradation

The paper would be incomplete if the discussion on the significance of the results obtained with the EkP system with respect to those attained with LVQ was missing. EkP with one, two and three prototypes per class selected eight times performed better than LVQ for which two codebooks were used. On the appendicitis, breast-cancer, german-credit and hepatitis the results were statistically insignificant. What concerns the statistical significance of the results obtained with EkP with respect to those attained with LVQ for which four codebooks were used, our system with one prototype per class outperformed the reference model five times (on credit-rating, Cleveland-heart, Hungarian-heart, heart-statlog and iris). The EkP system with two as well as with four prototypes per class outperformed LVQ

Java 1.6 on a notebook equipped with a Turion 64 ML-32 1.8GHz processor with a 2GB RAM memory.

additionally on pima-diabetes. In case of LVQ for which six codebooks were used, the results obtained with the *EkP* systems were statistically better than the ones obtained with LVQ four times (on credit-rating, Cleveland-heart, Hungarian-heart and heart-statlog).

The above discussion concerns the experiments with the *EkP* system for which 60 simplex points were used. Very similar trends have been observed in case of the experiments with 10 simplex points, however we will skip the details in order to avoid the paper becoming too dense.

## 4 Conclusions

It seems we were lucky to have created quite a fast prototype selection system despite of employing the simplex minimization routine which is usually expensive. The initial experiments indicate that the method may turn out to be competitive to other data pruning systems. In the preliminary calculations the method discussed in this paper have shown statistical insignificance of the generalization ability with respect to IB1 and sometimes turned out to be superior to the LVQ system ver. 1. We are however aware of the fact that the LVQ system exists in many variants and it would be possible to obtain better results with this method. However the *EkP* training times are longer than those of IB1 but the testing times performed on large problems are shorter. After all, one should remember about the general idea laying behind the selection of prototypes: once the instances are initially found (training sets are pruned), the tests on unseen samples which are usually frequently performed can be conducted much faster. Before the *EkP* system is not confronted with many other prototype selection algorithms and before further experiments with our method are not performed it will be hard to estimate a real value of our contribution to the pattern recognition field.

## References

- AHA, D. W., KIBLER, D. and ALBERT, M. K.. Instance-based learning algorithms. *Machine Learning*, 6, 37-66, 1991.
- BROWNLEE J. A java implementation of the SOM-LVQ PAK. <http://www.it.swin.edu.au/personal/jbrownlee/>
- DOBOSZ K. Statistical Significance Tests in Estimation of the Results Obtained with Various Systems that Learn. M.Sc. thesis, Nicolaus Copernicus University, Toruń, Poland, 2006 (In Polish)
- DUCH W., BLACHNIK M., Fuzzy rule-based systems derived from similarity to prototypes. *Lecture Notes in Computer Science*, Vol. 3316 (2004) 912-917
- DUCH W., GRUDZIŃSKI K., Prototype based rules - new way to understand the data. *IEEE International Joint Conference on Neural Networks*, Washington D.C. 14-18.07. 2001, pp. 1858-1863
- GROCHOWSKI, M.: Selecting Reference Vectors in Selected Methods for Classification. M.Sc. thesis, Nicolaus Copernicus University, Department of Applied Informatics, Toruń, Poland, (2003) (In Polish)

- GROCHOWSKI, M., JANKOWSKI N.: Comparison of Instance Selection Algorithms II: Results and Comments. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 580-585.
- GRUDZIŃSKI, K., DUCH, W.: SBL-PM: A Simple Algorithm for Selection of Reference Instances for Similarity-Based Methods. Intelligent Information Systems, Bystra, Poland (2000), in Advances in Soft Computing, Physica-Verlag (Springer), 99-108
- GRUDZIŃSKI, K.: SBL-PM-M: A System for Partial Memory Learning. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 586-591.
- HYNINEN, KANGAS, KOHONEN, LAAKSONNEN, TORKOLLA. LVQ\_PAK: The Learning Vector Quantization Program Package (1996)
- JANKOWSKI N., GROCHOWSKI, M.: Comparison of Instances Selection Algorithms I: Algorithms Survey. Artificial Intelligence and Soft Computing ICAISC 2004, in Lecture Notes in Artificial Intelligence (LNAI 3070), (Springer), 598-603.
- KOHONEN T., Self-Organizing Maps. Third ed. Berlin Heidelberg: Springer-Verlag; 2001. (Thomas S Huang; Teuvo Kohonen, and Manfred R. Schroeder. Springer Series in Information Sciences; 30).
- LAMPTON, M.: neldermead.java (<http://www.cea.berkeley.edu/mlampton/neldermead.java>)
- MALOOF, M., MICHALSKI, R. S.: Selecting Examples for Partial Memory Learning. Machine Learning, **41**, (2000) 27-52
- MERTZ, C. J., MURPHY, P. M.: UCI repository of machine learning databases. <http://www.ics.uci.edu/pub/machine-learning-data-bases>.
- NELDER, J. A., MEAD, R.: A simplex method for function minimization. Computer Journal **7** (1965), 308-313
- OJA M., KASKI S., KOHONEN T., Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum, Neural Computing Surveys, 3: 1-156., (2003).
- WILSON, D. R., MARTINEZ, T. R.: Instance Pruning Techniques. In Fisher, D.: Machine Learning: Proceedings of the Fourteenth International Conference. Morgan Kaufmann Publishers, San Francisco, CA, (1997), 404-417.
- WILSON, D. R., MARTINEZ, T. R.: Reduction Techniques for Instance-Based Learning Algorithms. Machine Learning, **38**, (2000), 257-286
- WITTEN I.H., FRANK E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, 2000.

