Navzkrižno testiranje simboličnih in konektivističnih pristopov strojnemu učenju na specializiranih bazah akutnega vnetja slepiča

Cross–testing Symbolic and Connectionist Machine Learning Approaches in Specialized Acute Appendicitis Databases

Avtor / Author

Milan Zorman[1,2], Sandi Pohorec[1], Bojan Butolen[1], Bojan Žlahtič[1], Peter Kokol[3,2]

Ustanova / Institute

[1]Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija; [2]Univerza v Mariboru, Medicinska fakulteta, Maribor, Slovenija; [3]Univerza v Mariboru, Fakulteta za zdravstvene vede, Maribor, Slovenija

[1]University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia; [2] University of Maribor, Faculty of Medicine, Maribor, Slovenia; [3]University of Maribor, Faculty of Health Sciences, Maribor, Slovenia

Ključne besede:
akutno vnetje slepiča, odločitvena drevesa, nevronske mreže, strojno učenje, navzkrižno testiranje, medicinska informatika

Key words:
acute appendicitis, decision trees, artificial neural networks, machine learning, cross–testing, medical informatics

Članek prispel / Received
12.01.2012

Članek sprejet / Accepted
26.05.2012

Naslov za dopisovanje / Correspondence

Red. prof. dr. Milan Zorman,
Univerza v Mariboru, Fakulteta
za elektrotehniko, računalništvo in
informatiko, Laboratorij za načrtovanje
sistemov, Smetanova ulica 17,
SI–2000 Maribor, Slovenija
Telefon +386 22207459;
E–pošta milan.zorman@uni–mb.si

Izvleček

Namen: *Na področju učenja z nadzorovanimi metodami strojnega učenja v medicini se pogosto srečujemo s pomanjkanjem učnih objektov, primernih za učenje klasifikatorjev. Najpogostejša vzroka za to sta pomanjkanje sredstev za izpeljavo raziskav in splošna redkost raziskovanega pojava, ki ne dovoljuje, da bi zbrali podatke o zadostnem številu primerkov.*

V tem prispevku predstavljamo rezultate raziskav, učenja klasifikatorjev in testiranj na podatkih, ki izhajajo iz treh baz na temo akutnega vnetja slepiča, ki imajo zelo podobno strukturo, različne velikosti in izhajajo iz različnih okolij.

Metode: *Odločili smo se za vzporedno izvedbo različice navzkrižnega testiranja (cross–testing) dveh vrst klasifikatorjev: odločitvenih dreves in umetnih nevronskih mrež. Naša razli-*

Abstract

Purpose: *In the world of learning with medical supervised machine approaches, we often face a lack of dataset objects suitable for training a classifier. Two of the most common reasons are the lack of funds to perform all of the required tests and dataset gathering, or simply the condition is too rare to collect a suitable number of cases. In this paper we present the results of a very rare opportunity to test and train classifiers on three acute appendicitis datasets with almost identical structures, but from different sources and of different sizes.*

Methods: *We performed a parallel variant of cross–testing of two types of classifiers (decision trees [DT] and artificial neural networks [ANN]). Our variant of cross–testing focuses on training the classifiers on one da-*

čica navzkrižnega učenja se posveča učenju klasifikatorja na eni bazi in preizkusu in testiranju na vseh preostalih testnih množicah, vključno s pripadajočo testno množico iz istega vira.Za primerjavo omenjenih treh Acute Abdominal Pain (AAP) baz podatkov različnih velikosti in virov smo izbrali osemnajst (18) parametrov anamneze in kliničnih preiskav ter odločitveni atribut/diagnozo akutno vnetje slepiča.

Rezultati: *Primerjavo rezultatov smo izvedli na osnovi splošne natančnosti, senzitivnosti in specifičnosti klasifikatorja ter uravnoteženosti slednjih dveh parametrov.*

Zaključek: *Dobljeni rezultati so presenetljivo odstopali od pričakovanj, saj faktorji, kot sta velikost učne množice in teoretična moč pristopa strojnega učenja niso pokazali pričakovanega vpliva.*

taset and testing all of the remaining available datasets, including the one derived from the same source, as the training set. To compare the three acute abdominal pain databases of different sizes and origins, we selected 18 parameters from patient medical histories, clinical examinations, and the decision attribute/diagnosis, acute appendicitis.

Results: *The comparison of results was based on overall accuracy, sensitivity, specificity, and balance.*

Conclusion: *The results we obtained were quite surprising, and factors such as dataset size and theoretical power of the methods did not prove to be as important as first expected.*

INTRODUCTION

In the world of learning with medical supervised machine approaches, we often face a lack of dataset objects suitable for training a classifier. Two of the most common reasons are the lack of funds to perform all of the required tests and dataset gathering, or simply the condition is too rare to collect a suitable number of cases. An insufficient number of dataset objects is manifested in a poor search space coverage, an unbalanced number of representatives for each class, or test and validation sets that are too small or non–existent. The final result is usually a classifier model that does not perform as well as it could, or is not thoroughly tested in a real world environment.

In this paper we present the results of a very rare opportunity to test and train classifiers on three acute appendicitis datasets with almost identical structures, but from different sources and of different sizes. We performed a parallel variant of cross–testing of two types of classifiers (decision trees [DT] and artificial neural networks [ANN]). Cross–testing builds on the idea of cross–training, which usually refers to a type of training in the domain of sports, in which an athlete trains for sports different from the discipline he/she competes in to improve overall performance. This approach is often used in athletics, mixed martial arts, professional training of rescue services, or even military special forces, and has been shown to be very successful (1).

Our variant of cross–testing focuses on training the classifiers on one dataset and testing all of the remaining available datasets, including the one derived from the same source, as the training set. The latter will serve as a reference for comparison of results derived from other test sets.

We expect to reveal interesting results of a direct comparison of two competent supervised machine learning approaches, decision trees, and neural networks to see if the myth that larger datasets contain more knowledge still stands.

In the following sections we will introduce the two approaches, the datasets, and finish with the results and conclusions.

SUPERVISED MACHINE LEARNING AND DECISION SUPPORT SYSTEMS

Decision support systems (DSS) assist physicians and are becoming a very important part of medical decision–making. DSS are based on different models, and the best of the DSS are providing an explanation together with an accurate, reliable, and quick response. Two of the most popular among machine–learning approaches are DT and ANN. Both DT and ANN have been successfully used in many medical decision–making applications for years because DT and ANN process data in ways that can be validated and interpreted. Where DT excels with transparent representation of acquired knowledge and rapid algorithms, what made one of the most often used symbolic machine learning approaches (2, 3), ANN persuade us with the superior classification power (4).

Acute appendicitis is a special problem in patients with acute abdominal pain (AAP) and presents one of the most difficult diagnostic challenges. The early and accurate diagnosis of acute appendicitis is still a difficult and demanding problem in clinical settings. Of major concern in patients with acute appendicitis is the perforation rate (up to 20%) and negative appendectomy rate (up to 30%) (5, 6). An important factor in the error rate is poor discrimination between acute appendicitis and other diseases that cause acute abdominal pain. This error rate is still high, despite considerable improvements in history–taking and clinical examination, computer–aided decision–support and special investigative modalities, such as ultrasonography.

Different types of automatic knowledge acquisition tools, such as DT (7) and ANN (8) have already evaluated databases with cases of acute abdominal pain. This clinical problem seems to be well–suited for supervised machine learning approaches because a standardized terminology has been defined. Agreed definitions, criteria, and minimum datasets have been promulgated by the World Organization of Gastroenterology (9).

Decision trees

DT is a typical representative of a supervised symbolic machine learning approach used for the classification of objects (10) in which patient data is presented with dataset objects represented in a form of attribute–value vectors. The structure is similar to a flowchart tree structure (see a sample DT in Figure 1); each internal node (rectangles in Figure 1) is a point of testing for an attribute value and based on it the classification continues throughout the left branch of the right sub–tree. Branches represent outcomes of node tests; each leaf node (circles in Figure 1) is a decision on the class to which an individual belongs (each leaf is class–labelled) (11).

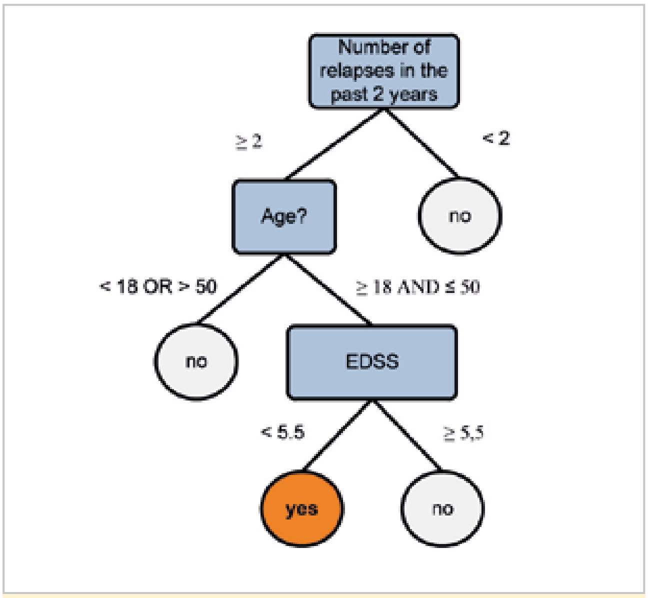


Figure 1. *A sample decision tree.*

Decision trees are used for classification in such a way that given an object, U, for which the classification class is unknown, the attribute values of U are tested against the nodes in the tree. The object (U) follows a path through the tree to a leaf node that assigns U with a class label. The DT is a well–established classifier because it allows exploratory knowledge discovery and does not require a detailed prior knowledge of the domain and setting of parameters. Additionally, DT can handle high–dimensional data. A traditional approach to the induction of a DT is with the use of the "divide and conquer" principle; specifically, if all of the objects belong to the same decision class then a tree is a single node, otherwise an attribute is selected and the set of objects is divided according to the splitting criterion of the selected attribute. The procedure starts in the root node as being most informative for the classification and is sequentially repeated for each of the branches going out of the first node. The first algorithm for DT induction was introduced by Quinlan in 1986 (12), the well–known iterative dichotomiser (ID)3. Quinlan also presented C4.5, a benchmark for newer supervised learning algorithms (13). Both ID3 and C4.5 use the information gain from a single attribute to build the DT. The attribute that adds the most information about the decision in the training set is selected first, then the next most informative of the remaining attributes follows until the leaf node is reached. Anomalies (or unbalanced data sets) can result in branches that are too complex or appear in multiple nodes (replication) in the tree. The solution to this is applying a tree-pruning procedure that shortens the tree, at the same time securing good accuracy.

Artificial Neural Networks

In our research we focused on multi-layer, feed-forward neural networks (sometimes called multi-layer perceptrons) and back-propagation learning methods. This type of neural network is widely used in various fields, and usually gives very good results (4).
The multi-layer architecture of ANN is most often used in practical applications. Each layer uses a linear combination function. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs.
Neural networks with no hidden layers should perform similarly to DTs; both approaches can divide a solution space by setting a hyper-plane in hyper-space (e.g., by setting a line in a 2-dimensional space to differentiate between different decision), (14, 15). Neural networks with one hidden layer are capable of limiting a more complex area (any open or closed convex area), where neural networks with two hidden layers can limit any area in hyper-space. Thus, more hidden layers in ANN result in a greater potential for solving complex problems.
Computational power and handling the noise in a training dataset are the greatest advantages of ANN; however, it takes a lot of time to build and train a network and gain insight into the accumulated knowledge, which is one of the important disadvantages when using ANN in domains that require knowledge representation suitable for manual inspection.

THE DATA SETS

To compare the three databases, we selected the following 18 parameters from history-taking and clinical examinations, which could be identified in all 3 databases:

- gender,
- age,
- progression of pain,
- duration of pain,
- type of pain,
- severity of pain,
- current location of pain,
- location of pain at the time of onset,
- previous similar complaints,
- previous abdominal surgery involving the appendix,
- distended abdomen,
- tenderness,
- severity of tenderness,
- movement of the abdominal wall,
- rigidity,
- rectal tenderness,
- rebound tenderness, and
- leukocytes.

All of the remaining parameters from the original datasets had a missing value rate > 10% and were therefore excluded from tests. Because we were focusing on the problem of separating the acute appendicitis diagnosis (class: "appendicitis") from other diseases that cause acute abdominal pain, these other diagnoses were labelled into one common class (class: "other diseases"). For a more detailed presentation of the methods used to collect the datasets and for an in-depth description of the parameters used, refer to previous publications (16, 17, 18).

AAP Databases

1. **AAP I (n = 1254):** This prospective clinical data base of AAP was built within the framework of a Concerted Action of the European Community (COMAC-BME-European Community Concerted Action on Objective Medical Decision Making in Patients with Acute Abdominal Pain) (16). The data was gathered in six surgical departments in Germany. All of the patients with acute abdominal pain < 1 week in duration were included in the study. A structured and standardized patient history and clinical examina-

tion were performed for every patient, and the data was gathered using a form based on the original abdominal pain chart of the World Organization of Gastroenterology (OMGE). The final diagnosis was based on operative findings, special investigations, and the course of the disease during the hospital stay. In cases of patients with non–specific abdominal pain, data from readmission and telephone interviews were used. The prevalence of appendicitis in this database was 16.8 % (n = 211).

2. AAP II (n = 2286): This prospective database was built during the German MEDWIS project A70 “Expert system for acute abdominal pain” 17). The data was derived from 14 centres in Germany. All of the patients with acute abdominal pain < 1 week in duration were included in the study. We used enhanced versions of the forms presented in the AAP I for data collection 16). The final diagnosis was based on the diagnosis at discharge. The prevalence of appendicitis in this database was 22.7 % (n = 519). This dataset contained special (more complicated) cases, thus patients were referred from general hospitals to university hospitals.
3. **AAP III (n = 4020):** This prospective database was built during a Concerted Action funded by the European Commission during the COPERNICUS programme no.: 555, “Information Technology for the Quality Assurance in Acute Abdominal Pain.” Data was collected in the 16 centres from central and eastern Europe. The data were collected in the same way as AAP II. Medical terminology was translated into 10 different languages, so that the participating centres could be provided with national versions of the software 18). The final diagnosis was based on the diagnosis at discharge. The prevalence of appendicitis in this database was 40.5% (n = 1628).

In all three databases, we additionally filtered out the cases for which > 90% of parameters were not known. As a result of this step, the number of cases in AAP I was reduced by 3 objects (from 1254 to 1251), the number of cases in AAP II was reduced by 7 cases (from 2286 to 2279), and the number of cases in AAP III remained the same.

To perform cross–testing, we prepared different variants of training and test sets to gain insight into the classifier’s performance.

Training and test sets

We decided not to use the training objects with > 10 missing values and tried to increase the quality of produced classifiers. We labelled the datasets which contained objects with < 10 missing values as the cleaned data sets.

During our previous tests we determined that the percentage of appendicitis cases in all three data sets was substantially < 50%. To improve the power to avoid bias of classifiers, we reduced the number of objects in the sets by removing the objects classified as ‘other diagnosis’ that had the greatest number of missing values, hereafter referred to as reduced datasets.

For each dataset, we built two training sets. For the first training set (labelled as Training set 50:50 in Tables 1–6) we used approximately two–thirds of the cleaned data set. The remaining one–third of the dataset was saved for testing purposes as the test set. Then, we reduced the two–third training set, so that it contained approximately the same number of appendicitis cases and cases marked as “other diagnosis.”

The second training set was the reduced data set (marked as Full set 50:50 in Tables 1–6) and was comprised of the original data set with an approximate ratio of 1:2 of objects classified as appendicitis cases and the other half was classified as “other diagnosis.”

The number of training objects in Training sets 50:50 was 274 for AAP I (137 classified as appendicitis and137 classified as other diagnosis), 763 for AAP II (363 classified as appendicitis and 400 classified as other diagnosis), and 2186 for AAP III (1086 classified as appendicitis and 1100 classified as other diagnosis).

The number of training objects in Full sets 50:50 was 422 for AAP I (211 classified as appendicitis and 211 classified as other diagnosis), 1119 for AAP II (519 classified as appendicitis and 600 classified as other diagnosis), and 3330 for AAP III (1628 classified as appendicitis and 1702 classified as other diagnosis).

The test sets were generated similar to the training sets. The difference was that we did not have the approximate 1:1 ratio of appendicitis cases and other diagnosis in the test set, which made the testing conditions even more realistic.

The first test set for each dataset (marked as Test set in Tables 1–6) was the remaining one-third of the cleaned dataset. The second test set was the same, which was actually a cleaned dataset (marked as Full set in Tables 1–6).

The number of test objects in 'test sets' was 414 for AAP I (74 classified as appendicitis and 340 classified as other diagnosis), 731 for AAP II (156 classified as appendicitis and 575 classified as other diagnosis), and 1340 for AAP III (542 classified as appendicitis and 798 classified as other diagnosis).

The number of test objects in 'full sets' was 1251 for AAP I (211 classified as appendicitis and 1040 classified as other diagnosis), 2279 for AAP II (519 classified as appendicitis and 1760 classified as other diagnosis), and 4020 for AAP III (1628 classified as appendicitis and 2329 classified as other diagnosis).

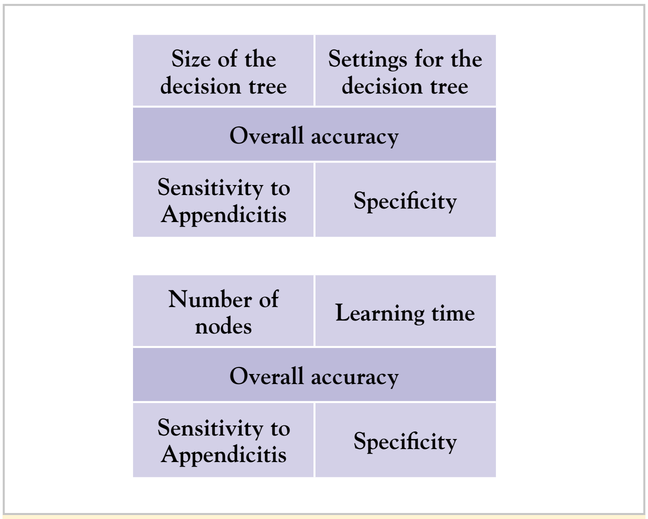
RESULTS

The results in this section were produced using the following two supervised machine approaches: DT, implemented in the MtDeciT2.1Gen environment (19); and ANN, implemented in the Weka environment (20). For each AAP dataset, we therefore created a separate table to collect the results of the cross-testing.

For each AAP data set, we built two types of DTs and ANNs (one for each version of the training set), and tested each of the classifiers on each possible test set, except the full set (for classifiers built on a training set) and the full and testing sets (for classifiers built on a reduced data set). The basis for the latter was that training sets contained a few objects that were in the test sets and which would diminish the objectivity of the results.

Each part of Tables 1–6 contains cells with data specific to each ML approach (see cell maps in Figure 2), as follows:

- number of nodes in the DT, settings for the DT (pre-pruning percentage, type of discretization technique; see (19) for an in-depth explanation), overall accuracy, sensitivity to appendicitis, and specificity; and
- number of nodes in the ANN, learning time in seconds, overall accuracy, sensitivity to appendicitis, and specificity.



Size of the decision tree	Settings for the decision tree
Overall accuracy	
Sensitivity to Appendicitis	Specificity

Number of nodes	Learning time
Overall accuracy	
Sensitivity to Appendicitis	Specificity

Figure 2. *Cell maps for testing results: left for decision trees, right for neural networks.*

In Table 1 the best results of the DTs built on the AAP I data set are shown and tested on each possible data set, except for the full AAP I set. It is noteworthy that the best accuracy was not achieved for the AAP I test set, but for the AAP III test and full sets.

Similar results can be observed with the ANN approach in Table 2. The overall accuracy is a bit lower than shown in Table 1, with the largest difference (8.01%) in tests using the AAP III full set. The balance between sensitivity and specificity was also a lower for ANN-generated classifiers.

In Table 3 the best results of the DTs built on the AAP II data set are shown. Using the training set, the best results on the AAP II test set were obtained. Somewhat unexpectedly, the best results for the DT built with the full AAP II set were achieved using the AAP I test and the full sets. In spite of that, no DTs built on AAP II could be described as clinically useful because in the majority of cases the best accuracy slightly exceeded 50%.

The overall accuracy of ANN (Table 4) was lower than with the DT shown in Table 3. The biggest difference in overall accuracy between Tables 3 and 4 was on the

Table 1. *Results of the comparison of the AAP I DT on different test sets*

		TEST SETS									
		SMALL (AAP I)		MEDIUM (AAP II)				LARGE (AAP III)			
		Test Set		Test Set		Full Set		Test Set		Full Set	
Small Training Set (AAP I)	Training Set–50:50	69 Nodes	25%DC40–2	69 Nodes	25%DC40–2	69 Nodes	25%DC40–2	69 Nodes	25%DC40–2	69 Nodes	25%DC40–2
		73.67%		56.77%		54.76%		75.52%		75.67%	
		74.32%	73.53%	42.95%	60.52%	39.11%	59.38%	65.87%	82.08%	64.99%	82.94%
	Full Set 50:50			23 Nodes	30%DC40–2	23 Nodes	30%DC40–2	23 Nodes	30%DC40–2	23 Nodes	30%DC40–2
				55.81%		52.87%		82.31%		81.99%	
				49.36%	57.57%	42.77%	55.85%	82.29%	82.33%	81.88%	82.07%

Table 2. *Results of the comparison of the AAP I ANN on different test sets*

		TEST SETS									
		SMALL (AAP I)		MEDIUM (AAP II)				LARGE (AAP III)			
		Test Set		Test Set		Full Set		Test Set		Full Set	
Small Training Set (AAP I)	Training Set–50:50	54	20.4s	54	20.44s	54	20.53s	54	20.47s	54	20.74s
		67.39%		54.17%		53.75%		71.42%		71.69%	
		30.57%	89.88%	20.66%	78.17%	21.21%	76.15%	65.03%	75.59%	65.16%	76.08%
	Full Set 50:50			54	32.05s	54	31.73s	54	31.31s	54	31.03s
				52.80%		52.70%		73.43%		73.98%	
				22.12%	79.34%	22.24%	76.81%	66.85%	78.05%	67.28%	78.81%

AAP I full set tests as it grew to 23.5% in favour of DT. The results of testing DTs, built on the AAP III training and full sets (Table 5), indicate that the AAP III data set is capable of providing more knowledge as the AAP II data set (Table 3). The highest accuracy of the DT built on the AAP III training and full sets (Table 3) matched the highest accuracy of the tree built on the AAP I training and full sets. But taking into account the average accuracy of the AAP I (Table 1) still gave slightly better results.

An overall comparison of the two approaches shows that the ANN displayed the lowest average accuracy and absolute difference in accuracy toward the DT results. The balance between sensitivity and specificity was also a bit lower for ANN, thus showing more bias toward one of the possible decisions.

A comparison of classification results using different data sets exposes the best average accuracy achieved by the DT, built on the small reduced data set AAP I (marked as 'Full set 50:50' in Table 1), followed closely by the DT built on the large AAP III data set (Table 5), and ANN built on the large AAP III data set (Table 6); considering the data set background, the sizes, and theoretical power of machine–learning approaches a surprising outcome.

Table 3. *Results of the comparison of the AAP II DT on different test sets*

		TEST SETS									
		SMALL (AAP I)				MEDIUM (AAP II)		LARGE (AAP III)			
		Test Set		Full Set		Test Set		Test Set		Full Set	
Medium Training Set (AAP II)	Training Set–50:50	464 Nodes	30%Q	464 Nodes	30%Q	47 Nodes	40%DC40–2	464 Nodes	30%Q	464 Nodes	30%Q
		42.75%		43.73%		52.39%		46.79%		43.73%	
		43.24%	42.65%	38.86%	44.71%	51.28%	52.70%	39.30%	51.88%	36.12%	48.91%
	Full Set 50:50	162 Nodes	40%DC40–2	162 Nodes	40%DC40–2			162 Nodes	40%DC40–2	162 Nodes	40%DC40–2
		68.84%		69.30%				56.79%		55.47%	
		54.05%	72.06%	44.08%	74.42%			32.84%	73.06%	31.27%	71.95%

Table 4. *Results of the comparison of the AAP II ANN on different test sets*

		TEST SETS									
		SMALL (AAP I)				MEDIUM (AAP II)		LARGE (AAP III)			
		Test Set		Full Set		Test Set		Test Set		Full Set	
Medium Training Set (AAP II)	Training Set–50:50	54	56.38s	54	56.2s	54	56.29s	54	56.39s	54	56.34s
		39.86%		39.89%		46.24%		45.37%		43.66%	
		17.71%	81.82%	16.7281%	82.88%	23.84%	82.73%	35.61%	54.85%	34.07%	53.14%
	Full Set 50:50	54	82.73s	54	82.28s			54	82.62s	54	82.81s
		46.14%		45.80%				54.85%		55.12%	
		15.67%	79.70%	16.21%	82.32%			44.63%	62.82%	44.76%	62.56%

DISCUSSION AND CONCLUSIONS

By knowing the background and methods used to collect the presented datasets, we did not expect the classifiers built on a medium data set from AAP II to perform so poorly. Somewhat poorer results were achieved with the AAP II test set using the DT and ANN that were built on the training sets of the AAP II data set.

The only reasonable explanation is that the AAP II dataset contains a large number of special cases which affected the applied machine learning approaches and cannot exploit the training objects as expected. The overall accuracy of the remaining comparisons between the AAP I and AAP III datasets was so high that some of those classifiers are of practical use to clinicians.

The accuracy we achieved during our experiments with DT and ANN was substantially higher than the accuracy as previously reported on approaches, such as neural networks (21) or case–based reasoning (22). The results presented herein show that cross–testing paid off, even in putting to test all combinations of training/testing sets. Based on the literature and the former research in other clinical domains, we had expected a better performance from the ANN results in

Table 5. *Results of the comparison of the AAP III DT on different test sets*

		TEST SETS									
		SMALL (AAP I)				MEDIUM (AAP II)				LARGE (AAP III)	
		Test Set		Full Set		Test Set		Full Set		Test Set	
Large Training Set (AAP III)	Training Set–50:50	48 Nodes	40%DC40–2	48 Nodes	40%DC40–2	48 Nodes	40%DC40–2	48 Nodes	40%DC40–2	48 Nodes	40%DC40–2
		64.73%		63.39%		53.21%		50.59%		83.81%	
		86.49%	60.0%	88.63%	58.27%	52.56%	53.39%	45.86%	51.99%	89.48%	79.95%
	Full Set 50:50	43 Nodes	40%DC40–2	43 Nodes	40%DC40–2	43 Nodes	40%DC40–2	43 Nodes	40%DC40–2		
		66.67%		65.87%		54.17%		51.73%			
		86.49%	62.35%	88.15%	61.35%	51.92%	54.78%	45.09%	53.69%		

Table 6. *Results of the comparison of the AAP III ANN on different test sets*

		TEST SETS									
		SMALL (AAP I)				MEDIUM (AAP II)				LARGE (AAP III)	
		Test Set		Full Set		Test Set		Full Set		Test Set	
Large Training Set (AAP III)	Training Set–50:50	54	162.56s	54	161.9s	54	161.24s	54	161.44s	54	161.59s
		59.66%		59.31%		49.79%		48.05%		81.87%	
		27.75%	92.20%	26.35%	92.75%	21.0958%	78.42%	20.60%	75.09%	72.55%	90.99%
	Full Set 50:50	54	244.75s	54	244.78s	54	244.89s	54	244.52s		
		63.77%		63.07%		55.27%		54.71%			
		30.41%	93.18%	28.40%	93.13	21.40%	78.70%	20.59%	75.84%		

terms of overall accuracy, sensitivity, and specificity. As shown in Tables 1–6, this was not the case.

Gathering data from different types of sources can substantially influence the performance of classifiers, even though the methods for data gathering were almost the same. We have also shown that larger training sets do not necessary guarantee a higher accuracy in comparison to smaller training sets. The results might still be improved, i.e., the quality of the extracted knowledge through sampling the training objects (cross–training) from all three training sets simultaneously. Nevertheless, of a positive impact for the clinical use is the fact that even a smaller dataset could provide valid, meaningful knowledge. The research presented herein has suggested that the following variables had contributed to the correct classification, as follows: tenderness (location); rectal tenderness; duration of pain; type of pain; and leucocytes.

ACKNOWLEDGEMENTS

Part of the work was supported in the framework of a bilateral scientific–technological cooperation between Slovenia (Ministry of Science and Technology, Project No.: L2–1640–0796–99) and Germany (Deutsches Zentrum für Luft– und Raumfahrte.V., Project No.: SVN99/022).

REFERENCES

1. Quinn E: Cross Training Improves Fitness and Reduces Injury. About.com Guide Updated October 28, 2008, http://sportsmedicine.about.com/od/tipsandtricks/a/Cross_Training.htm [last accessed 10.01.2012].
2. Klercker T AF: Effect of pruning of a decision–tree for the ear, nose and throat realm in primary health care based on case–notes. Journal of Medical Systems, 1996; 20(4): 215–26.
3. Zorman M, Podgorelec V, Kokol P, Peterson M, Lane J: Decision tree's induction strategies evaluated on a hard real world problem. In: 13th IEEE symposium on computer–based medical systems 22–24 June 2000, Houston, Texas, USA: Los Alamitos, IEEE Computer society 2000: 19–24.
4. Hornik K: Approximation capabilities of multilayer feedforward networks. Neural Networks 1991; 4(2): 251–57.
5. Andersson RE, Hungander A, Thulin JG: Diagnostic accuracy and perforation rate in appendicitis: association with age and sex of the patient and with appendectomy rate. European Journal of Surgery 1992; 158: 37–41.
6. Blind PJ, Dahlgren ST: The continuing challenge of the negative appendix. Acta Chir. Scand. 1986; 152: 623–27.
7. Ohmann C, Moustakis V, Yang Q, Lang K: Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. Artificial Intelligence in Medicine 8 1996; 23–36.
8. Pesonen E, Eskelinen M, Juhola M: Comparison of different neural network algorithms in the diagnosis of acute appendicitis, International Journal of Bio–Medical Computing, 1996; 40: 227–33.
9. deDombal FT: Diagnosis of Acute Abdominal Pain, Churchhill Livingstone, Edinburgh, (1991) 105–6.
10. Podgorelec V, Zorman M: Decision Trees: Encyclopedia of complexity and systems science, New York: Springer. 2009; 2:1826–45,
11. Han J: Data Mining: Concepts and Techniques San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2005.
12. Quinlan JR: Induction of decision trees, Machine Learning. 1986; 1:81–106.
13. Quinlan JR: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, 1993.
14. Russel SJ, Norvig P et al.: Artificial intelligence: a modern approach. Englewood cliffs, Prentice–Hall 1995; 525–62.
15. Rojas R: Neural Networks: A Systematic Introduction, Springer Verlag, Berlin, 1996.
16. de Dombal FT, de Baere H, van Elk PJ, Fingerhut A, Henriques J, Lavelle SM et al.: Objective medical decision making in acute abdoinal pain, In: Benken JEW and ThevinV (Eds.): Advances in Biomedical Engineering, IOS Press, 1993; 65–87.
17. Ohmann C, Platen C, Belenky G, Franke C, Otterbeck R, Lang K et al.: Expertensystem zur Unterstützung von Diagnosestellung und Therapiewahl bei akuten Bauchschmerzen. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 1995; 26(3): 262–74.
18. Ohmann C, Eich HP, Sippel H: A data dictionary approach to multilingual documentation and decision support for the diagnosis of acute abdominal pain (COPERNICUS 555, An European Concerted Action). Medinfo 1998; 9(1): 462–66.
19. Zorman M, Hleb Š, Šprogar M: Advanced tool for building decision trees MtDeciT 2.0. In: Kokol P, Welzer–Družovec T, Arabnia HR. (Eds.). International conference on artificial intelligence, June 28 – July 1, 1999, Las Vegas, Nevada, USA. Las Vegas: CSREA, 1999, book. 1: 315–18.
20. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
21. Pesonen E, Ohmann C, Eskelinen M, Juhola M: Increasing the accuracy of the acute appendicitis of a LVQ neural network by the use of larger neighbourhoods. Methods of Information in Medicine 37(1) 1996; 59–63.
22. Puppe B, Ohmann C, Goos K, Puppe F, Mootz O: Evaluating four diagnostic methods with acute abdominal pain cases. Methods of Information in Medicine 34 1995; 361–68.