ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2009. No. 2(90)*
ELEKTRONIKA IR ELEKTROTECHNIKA

T125 *AUTOMATION, ROBOTICS*
AUTOMATIZAVIMAS, ROBOTECHNIKA

Methodology of Adaptation of Data Mining Methods for Medical Decision Support: Case Study

V. Špečkauskienė, A. Lukoševičius
Biomedical Engineering Institute at Kaunas University of Technology,
Studentų str. 65 Kaunas, LT−51359 Lithuania, phone: +370 37 407119, e-mail: vita.speckauskiene@kmu.lt;

Introduction

Data mining is a problem solving technique, which analyzes the data already stored in the data base. It is a process of discovering, classifying and finding patterns in data. The classification of data helps to make a decision in different types of problems. Machine Learning, Statistical and Neural Network algorithms are applied for efficient data mining. The problem is to find algorithms suitable to apply in order to discover relationships between data attributes and make predictions that could be useful for decision support. In addition, there are several well known medical data problems, such as incorrect and sparse information and temporal data [1]. To deal with these problems there are supervised and unsupervised learning algorithms. The latter functions with missing values in the data. Overall problem of data mining is common in analyzing the literature. For example [2] and [3] analyze different types of algorithms, while [4] and [5] concentrate on Machine Learning algorithms.

The aim of this paper is to elaborate and test a method of data mining algorithms and the adjustment of these algorithms for decision support. In this case we apply the steps stated in our method on data mining algorithms and eye health screening data.

Method

Data is the most important aspect of efficient data mining. These techniques are applied on data and their performance is highly dependant on specificity and quality of data. The specificity of data is related to it's form: text (nominal), numerical, imagery (multimedia). The important aspect of data is ability to classify them, for example numerical values classified into significant groups; in that case we have a more important (obvious for the algorithm) attribute. Multimedia is analyzed with different algorithms than text or numbers, so before using, multimedia should be parameterized. So the first step is data analysis and data set (DS) forming.

The quality of data is more related to algorithms and the separation of algorithms suitable for the DS. Data can contain missing values; it means that not all attribute values are known. There are algorithms that operate with such data, others don't. There are also algorithms that only operate on nominal data, others on weighted. So the second step is to collect algorithms according to data specifics.

The easiest way to collect algorithms is to choose a data mining environment. Such distributed under GPL software is available online, e.g. "Orange" [6], "Weka" [7] and others. So, it is easier to discard unsatisfactory algorithms if the data specific is known.

In summarizing we can separated the following methodological steps:

1. Collecting and getting acquainted with a number of classification algorithms (e.g. data mining environment).
2. Reviewing the data set (e.g. a part of a patient health records).
3. Separating appropriate algorithms suitable for the DS.
4. Testing the full data set on selected number of classification algorithms, containing their default parameter values.
5. Selecting the best algorithms to use for further experiments.
6. Training the selected algorithms on reduced data set, by removing the attributes that appeared to be uninformative in building and visualizing the decision trees. Uninformative attributes are the ones that don't appear in the tree, or are at the leaf nodes.
7. Modifying algorithms' default parameter values. Using the optimal data set formed for each algorithm of the most useful data identified in step 6.
8. Evaluating the results.
9. Randomizing the data set.
10. Performing steps 6 and 7 on randomized data set.
11. Evaluating and comparing results as well as algorithms performance.

These steps can be performed using any data mining environment and any data set as well as any class attribute of the data set.

Method evaluation

The data set used for experiments was collected during eye health screening examinations in Eye Clinic of Kaunas University of Medicine. It contains 1140 instances of 12 category attributes (1140x12 matrix). The attributes contain certain ophthalmologic data. All the attributes were nominal (maximum 6 categories in attribute) and didn't contain any missing values. For experiments we selected two class attributes, one more obvious, and other less obvious. The class attribute could be any of 12 attributes in the DS, but we selected the ones that raise more and less obvious medical problem. The data set was used for both – training (66% of the data) and testing (34% of the data), as recommended in [7].

Classification was performed using WEKA data mining environment [7]. We collected 54 algorithms suitable for our DS. The algorithms were of six method groups. The methods are described in Table 1.

Table 1. Description of methods used for experiments

Method group	Number of algorithms in the group	Description
Bayesian Classifiers	5	Uses the technique of calculating probabilistic distributions.
Decision Trees	9	Based on finding the root node and splitting the attributes down the tree, until the final (decision) leaf is reached; usually using divide-and-conquer approach.
Rules	9	Uses the approach of covering, because at each stage a rule is identified so that it "covers" some of the instances in the data set. The rules can be of different types, e.g. classification, association, rules with exceptions, etc.
Linear Models	7	Based on linear regression, logistic regression, linear classification using the perceptron and other modifications.
Instance-based Learning	5	A distance function is used to calculate which member of the training set is closest to an unknown test instance (nearest neighbor approach).
Metalearning Algorithms	19	Use classifiers (one or two of the already mentioned groups) together with special schemes for reducing iterations, filtering data, performing regression and optimizing the data set.
Total: 6 groups	Total: 54 algorithms	

As already mentioned, the algorithms had to operate on two different class attributes of the DS. For the evaluation of algorithms the decisive parameters were sensitivity (%) and specificity (%). Those values came from the Classifier output window provided in WEKA.

Firstly, steps 4-10 were performed on more obvious class attribute. Secondly, steps 6-7 were performed on less obvious class attribute.

In step 7 we alternated the algorithm parameter values and noted the influences to sensitivity and specificity. In step 10 we noted the influence of randomization by comparing sensitivity and specificity. Lastly, we compared the results supplied on each question. Measured results are presented in the next section.

Results

Experiments of the third step showed very low specificity rates. Only 7 out of 54 algorithms produced specificity higher than 50%. However the sensitivity values were a lot higher – almost all algorithms reached more than 90%. We selected 8 algorithms that reached specificity more than 40%. The algorithms were of 5 method groups. The best results reached Decision Trees (NBTree, ADTree, REPTree), Metalearning Algorithms (LogitBoost, RandomCommittee), Linear Model (VotedPerceptron), Bayesian Classifier (AODE) and Instance-based Learning (LWL) algorithms. The results are visualized in Fig. 1.

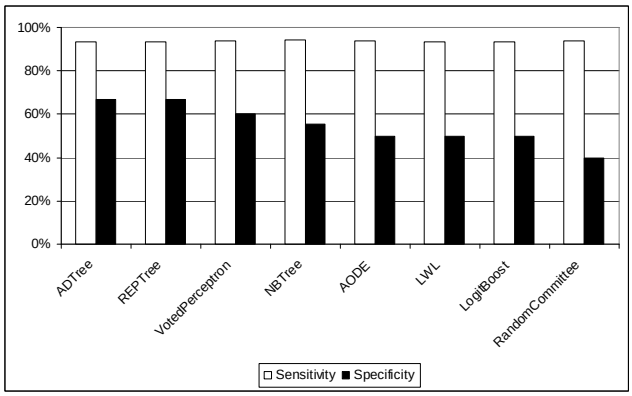

Fig. 1. The results of third step experiment: sensitivity (%) and specificity (%) of best performed algorithms

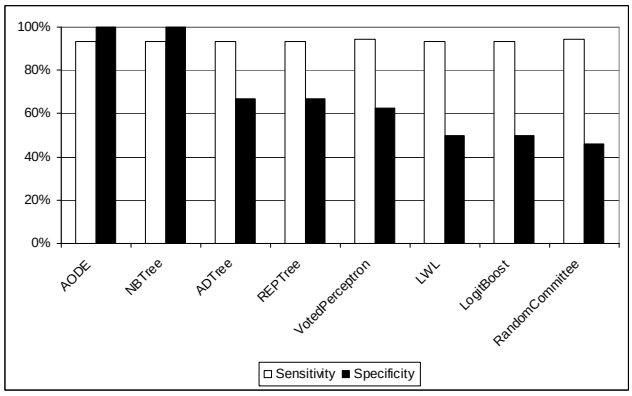

Fig. 2. The results of sixth step experiment: sensitivity (%) and specificity (%) of best performed algorithms after iteratively removing the most uninformative parameters

In the sixth step we dealt with the selected 8 algorithms. By visualizing the trees built by Decision Tree algorithms, namely ADTree and NBTree, we started to iterate the algorithms by removing the most uninformative parameters in the data set. The algorithms reached different

sensitivity and specificity rates. The best were of AODE and NBTree (sensitivity 94%, specificity 100%). The results are presented in Fig. 2.

The best performance of the algorithms is achieved with the different number of parameters in the data set, which was used in step 7. Randomization of a data set was performed using a randomize function implemented in WEKA. Once randomized, the data set was used to perform step 10. The results and explanations of steps 7-10 are detailed in Table 2.

Table 2. Modification of algorithms and randomization of DS: influences to sensitivity (%) and specificity (%)

Algorithm name	**Reached sensitivity (%) and specificity (%); parameters influence, comparison of results**	
	Initial DS	**Random DS**
AODE (Bayesian classifier)	94%, 100%	94%, 100%
	Don't have any parameter values to be changed, but different number of attributes were in initial and randomized DS to reach presented results.	
NBTree (decision tree)	94%, 100%	95%, 63%
	Don't have any parameter values to be changed, but different number of attributes were in initial and randomized DS to reach presented results.	
ADTree (decision tree)	94%, 100%	94%, 100%
	Number of boosting iterations were changed to 6 to reach 100% of specificity in initial DS. In randomized DS best results were reached without modifying this parameter.	
LogitBoost (metalearning algorithm)	94%, 100%	94%, 100%
	Using resampling came out with better results than reweighting, raising the specificity to 100% in initial DS. In randomized DS best results were reached without modifying this parameter.	
REPTree (decision tree)	94%, 67%	94%, 100%
	Better results were not reached in initial DS, while in randomized DS the best results were reached without pruning the decision tree, raising the specificity to 100%.	
LWL (instance-based learning algorithm)	94%, 67%	94%, 100%
	Default Linear and Tricube functions performed better, rising sensitivity to 94%, specificity to 67% in initial DS. In randomized DS only Tricube function raised sensitivity to 94%, specificity to 100%.	
VotedPerceptron (linear model)	94%, 63%	94%, 100%
	Better results were not reached in initial DS, but in randomized DS better specificity (100%) was reached without modifying any parameter.	
RandomCommittee (metalearning algorithm)	94%, 46%	94%, 50%

Table 2 (Continuation)

	Better results were not reached in initial DS, but in randomized DS better specificity (100%) was reached without modifying any parameter.

In short, we can see that the highest results reached for the used data set was 94% sensitivity, and 100% specificity, being very high in overall data mining problem. Randomizing data didn't significantly change the results. Although, it influenced the performance of a few algorithms.

Lastly, we selected less obvious class attribute from the DS and performed calculations by removing the uninformative attributes (step 6) and modified algorithms' default parameter values (step 7). The results are assembled in Table 3.

Table 3. Results on less obvious class attribute

Algorithm name	**Reached sensitivity(%) and specificity (%)**	**Modifications and influences**
ADTree (decision tree)	60%, 72%	Number of boosting iterations again set to 6 reached 60% of sensitivity and 72% of specificity with full DS.
VotedPerceptron (linear model)	58%, 73%	Best result was reached with full DS; modification of algorithm's parameters didn't attain better results.
AODE (Bayesian classifier)	50%, 72%	Best result was reached with full DS.
LogitBoost (metalearning algorithm)	50%, 72%	Using resampling with full DS came out with better results than reweighting, raising the sensitivity to 50%.
NBTree (decision tree)	50%, 72%	Best result was reached with full DS.
REPTree (decision tree)	39%, 73%	The minimum total weight of the instances in a leaf set between 0,5 and 1 raised sensitivity to 39% and specificity to 73% with full DS.
RandomCommittee (metalearning algorithm)	36%, 73%	Number of iterations set to 5 reached 36% of sensitivity and 72% of specificity with full DS.
LWL (instance-based learning algorithm)	0%, 72%	Performance of this algorithm was poor. Any changes didn't cause better results.

Obviously the algorithms reached a lot worse results than in the first case. The influence of number of attributes being in the DS was less important than in the first case. Although, modification of algorithm parameters had a big influence. The positions of algorithms with higher results are also slightly different; performance of VotedPerceptron is a lot better in this case, while of NBTree – a lot worse. Still the best performance was of Decision Tree algorithm ADTree, Bayesian classifier AODE and Metalearning algorithm LogitBoost.

Discussion and conclusions

This paper presents a methodology of selection, adjustment and application of data mining algorithms for decision support. The advantages of the proposed method lies in the wide scope of classification algorithms taken into account (54) which covers practically all known algorithms and also in the ability to estimate all improvements of algorithms and primary parameters (attributes) by calculation of quantifiable results of classification. It is not related to data nor to data mining techniques. The processes to be performed are not strictly related to one another and can be performed if needed. This gives a chance to objectively compare this method with other similar methods in future. Also, using this method, it is possible to indicate meaningful algorithm parameters and evaluate them quantifiably.

In summary, of the results of method evaluation and analyzed algorithms the best outcome was gained using Decision Tree algorithm ADTree, Bayesian classifier AODE and Metalearning algorithm LogitBoost.

Acknowledgment

Authors would like to acknowledge the support of the Lithuanian State Science and Studies Foundation for funding of the research project “Info Sveikata” (“Info Health”), reg. No. B-07019

References

1. **Serrano J. I., Tomeckova M., Zvarova J.** Machine Learning Methods for Knowledge Discovery in Medical Data on Atherosclerosis // EJBI 2006. – No. 1. – P. 6–33.
2. **Su J. L., Wu G. Z., Chao I. P.** The Approach of Data Mining Methods for Medical Database // Engineering in Medicine and Biology Society 2001. Proceedings of the 23rd Annual International Conference of the IEEE. – Vol. 4. – P. 3824–3826.
3. **Burn-Thornton K. E., Lazzarini A.** A Toolkit to Aid Clinical Decision Support // Information Technology Applications in Biomedicine 2003. 4th International IEEE EMBS Special Topic Conference. – P. 242–245.
4. **Mertik M., Kokol P., Zalar B.** Gaining Features in Medicine Using Various Data-Mining Techniques // Computational Cybernetics ICCC 2005, IEEE 3rd International Conference. – 2005. – P. 21–24.
5. **Jegelevičius D., Lukoševičius A., Paunksnis A., Barzdžiukas V.** Application of Data Mining Technique for Diagnosis of Posterior Uveal Melanoma // Informatica, Lith. Acad. Sci. – 2002. – No. 13(4). – P. 455–464.
6. **Demsar J., Zupan B., Leban G.** Orange: From Experimental Machine Learning to Interactive Data Mining // White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana. – 2004.
7. **Ian H. W., Eibe F.** Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition. – San Francisco: Morgan Kaufmann, 2005.

Received 2008 02 11

V. Špečkauskienė, A. Lukoševičius. Methodology of Adaptation of Data Mining Methods for Medical Decision Support: Case Study // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009. – No. 2(90). – P. 25–28.

Data mining algorithms are used in various fields as a tool for answering (suggesting) particular questions. Medicine is one of the fields, which widely uses data mining techniques. In this research they are analyzed and applied in the domain of eye health. Aim: elaborate and test a method of data mining algorithms and adjustment of these algorithms for decision support. In this case we apply the steps stated in our method on data mining algorithms and eye health screening data. Paper presents the results of the testing of methodology, covering operation of algorithms’ specificity (%) and sensitivity (%). Highest reached sensitivity is 94%, specificity – 100%. Ill 2, bibl. 7 (in English; summaries in English, Russian and Lithuanian).

Б. Шпечкаускене, А. Лукошявичус. Методология применения методов добычи и анализа данных для организации поддержки принятия решений в медицине: исследование случаев // Электроника и электротехника. – Каунас: Технология, 2009. – № 2(90). – P. 25–28.

Алгоритмы добычи данных приминяются в различных областях как вспомогательное пособие ответить (посоветовать) на некоторые вопросы. Медицина является одной из областей, в которой широко применяются технологии добычи данных. В настоящем исследовании они анализируются и применяются в области глазных заболеваний. Цель: усовершенствовать и испытать метод добычи данных, предназначенных для применения в медицинских решениях. Для достижения данной цели в методике перечислены шаги применимы в алгоритмах добычи данных и данных в области глазных заболеваний. В публикации представлены результаты испытаний методики, каторая охватывает чувствительность дейстия алгоритмов (%) и специфичность (%). Получена наиболшая чувствительность – 94 %, специфичность – 100 %. Il. 2, библ. 7 (на английском языке; рефераты на английском, русском и литовском яз.).

V. Špečkauskienė, A. Lukoševičius. Duomenų gavybos ir analizės metodų taikymo medicininiams sprendimams rengti metodologija: atvejų tyrimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2009. – Nr. 2(90). – P. 25–28.

Duomenų gavybos ir analizės algoritmai naudojami įvairiose srityse kaip pagalbinė priemonė atsakyti į tam tikrus klausimus (patarti). Medicina – viena iš sričių, kurioje plačiai naudojamos duomenų gavybos ir analizės technologijos. Šiame tyrime jos analizuotos ir taikytos akių ligoms. Tikslas – ištobulinti ir išbandyti duomenų gavybos ir analizės metodą, skirtą taikyti medicininiams sprendimams rengti. Šiam tikslui pasiekti metodikoje išvardyti žingsniai taikomi duomenų gavybos ir analizės algoritmams ir akių ligų duomenims. Straipsnyje pateikiami metodikos išbandymo rezultatai, apimantys algoritmų veikimo jautrumą (%) ir specifiškumą (%). Didžiausias pasiektas jautrumas 94 %, specifiškumas 100 %. Il. 2, bibl. 7 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).