# Využití strojového učení pro predikci mortality

**Jiří Klema**
**Gerstnerova laboratoř pro inteligentní rozhodování a řízení, FEL ČVUT, Praha**

## Souhrn

Zdravotnictví je jednou z nejčastějších oblastí aplikace umělé inteligence a metod strojového učení. Tento článek se zaměřuje na predikci výsledku kardiovaskulární intervence na základě pacientovy anamnézy a jeho předoperačního stavu. Ve své úvodní části nejprve obecně definuje úlohu prediktivního získávání znalostí z dat a představuje dílčí metody vhodné pro aplikaci na sledovanou úlohu. V druhé praktické části pak prezentuje a zdůvodňuje výsledky predikce mortality dosažené nad Sloučeným národním registrem kardiovaskulárních intervencí (NZR - Databáze kardiovaskulárních intervencí) spravovaným střediskem MEDICON. Dosažené výsledky porovnává s výsledky bayesovského pravděpodobnostního modelu implementovaného v současné době používaném informačním systému PATS (Patient Analysis and Tracking System).

## Klíčová slova:

**Získávání znalostí z dat, strojové učení, rozhodovací a regresní strom, neuronová síť, případové usuzování, ROC plocha, registr srdečních operací, predikce mortality**

## Abstract

### Utilization of Machine Learning for Mortality Prediction

This article presents and evaluates general possibilities of machine learning application aimed to predict a raw result of patient operation (dead – living, nearmiss+ yes- no) on the basis of the patient anamnesis and his/her pre-operative state. It introduces briefly a task of predictive data mining as well as applied machine learning methods. At the same time it presents mortality prediction results when applied to Merged Czech National Cardiac Registry (MCZR) administered by MEDICON Center. The reached results are compared to results obtained by a Bayesian model implemented within a currently used informational system PATS (Patient Analysis and Tracking System).

## Keywords:

**Data mining, machine learning, decision and regression tree, neural network, instance-based learning, ROC area, cardiac registry, mortality prediction**

## Získávání znalostí z dat

Zdravotní registry obsahují velké množství údajů, navíc jsou neustále doplňovány o nové záznamy. Z důvodu přístupu k těmto údajům, jejich zabezpečení a následného zpracování je evidentní, že tyto registry musí být provozovány v rámci informačního systému. Předmětem tohoto článku samozřejmě není diskuse tohoto obecně známého faktu, ale studie jedné z možností obohacení stávajících informačních systémů o nejnovější teoretické možnosti automatizovaného získávání znalostí z dat.

Automatizované získávání znalostí z dat (Data Mining) je nově se rozvíjející vědní obor, definovaný jako kooperativní snaha lidí a počítačů o automatické vyhledávání užitečných informací v rozsáhlých datech [1]. Problematika je chápána v celé své šíři, velká pozornost je věnována návrhu databází, popisu problému a stanovení cílů. Toto jsou typické úkoly lidí. Počítačové algoritmy poté analyzují nahromaděná data a vyhledávají závislosti související s cíli definovanými lidmi.

Data mining ve své podstatě integruje zkušenosti získané ze tří dříve volně spojovaných oborů:

- databází a datových skladů,
- matematické statistiky,
- strojového učení.

Techniky získávání znalostí z dat se zaměřují na dvě problémové oblasti. První z nich je predikce, druhou pak vlastní extrakce znalostí. V rámci využití údajů nahromaděných ve zdravotních registrech lze aplikovat algoritmy obou zmíněných problémových oblastí. Algoritmy patřící do první prediktivní oblasti lze využít přímo ke konstrukci prediktivních modelů sloužících například při posuzování vhodnosti elektivních

intervencí, algoritmy druhé skupiny lze aplikovat například jako podporu při vytváření znalostní báze expertního systému.

## Přehled prediktivních metod

Často používanou třídou algoritmů jsou *rozhodovací stromy* [7]. Jedná se o učící se algoritmy, které vytvářejí rozhodovací hierarchické stromové struktury na základě klasifikované množiny trénovacích příkladů. Jejich základní myšlenka je založena na strategii "rozděl a panuj". Na každé rozhodovací úrovni volíme takový atribut (veličinu), který nejlépe odděluje jednotlivé klasifikační třídy. Aplikací tohoto kritéria se snažíme o to, aby vzniklý strom byl co nejmenší. Praktické zkušenosti totiž ukazují, že jednodušší stromy vykazují větší klasifikační přesnost. V případě numerické predikce, kdy namísto přiřazení příkladu ke klasifikační třídě (např. úspěšná operace – ano, úspěšná operace – ne) předpovídáme přesnou hodnotu (např. pravděpodobnost úspěšné operace – 87%), hovoříme o *regresních stromech*.
Dalším populárním algoritmem jsou *neuronové sítě.* Původním cílem výzkumu neuronových sítí byla snaha pochopit a modelovat, jakým způsobem myslíme a jakým způsobem funguje lidský mozek. Právě neurofyziologické poznatky umožnily vytvořit zjednodušené matematické modely, které se dají využít při řešení praktických úloh z umělé inteligence. Základem matematického modelu neuronové sítě je formální neuron, který je převodem funkce neurofyziologického neuronu do matematické řeči. Neuronové sítě se skládají z formálních neuronů, které jsou vzájemně propojeny tak, že výstup neuronu je vstupem obecně více neuronů. Počet neuronů a jejich vzájemná propojení v síti určuje architekturu (topologii) sítě. Druhým základním rysem neuronových sítí je jejich schoponost učení. Existuje celá řada učících algoritmů pro různé modely neuronových sítí. Nejznámějším a nejpoužívanějším modelem neuronových sítí je vícevrstvá neuronová síť s učícím se algoritmem zpětného šíření (backpropagation).
Podstata *případového usuzování* spočívá v tom, že řešitel přistupuje k problému tak, že se snaží najít jeho podobnost s problémy řešenými a úspěšně vyřešenými dříve. Nehledá tak pro nový problém nové unikátní řešení, ale snaží se jej odvodit z modifikací řešení úspěšných v minulosti. Z definice plyne, že se jedná o postup analogický odvozený z jednání vlastního lidem. Podobně jako v případě lidí pojem strojového případového usuzování zahrnuje celou řadu přesněji definovatelných postupů: modifikace dřívějších řešení, jejich kombinace, vysvětlení nově vzniklých situací, kritické zhodnocení a revize nově vzniklých řešení či vytváření nových příkladů na základě příkladů dřívějších. V případě diagnostických úloh může být redukován na *metodu nejbližších sousedů*.
Výše uvedené metody reprezentují nejčastěji používané metody strojového učení. Dobrých výsledků lze dosáhnout i pomocí metod statistických, například *lineárního regresního modelu*. Dalšího zlepšení výsledků lineárních regresních modelů, které jsou omezeny pouze na hledání lineárních kombinací vstupních atributů, lze dosáhnout aplikací nelineární regrese, popřípadě zavedením pojmu lokality. Lokalita použitého regresního postupu spočívá v předvýběru trénovacích příkladů použitých pro vytváření regresního modelu.
Další známou statistickou metodou učení je pravděpodobnostní *bayesovské učení*. Je založeno na Bayesově vztahu a jeho použití je vhodné zejména v případě zpracování znalostí zatížených nejistotou. Tzv. naivní Bayesův algoritmus je založen na změně pravděpodobnostního ohodnocení cílových hypotéz na základě hodnot jednotlivých atributů. Jedná se o jednoduchou metodu, která je ovšem zatížena chybou pokud jsou jednotlivé atributy vzájemně závislé. Umělá inteligence rozvinula výše zmíněnou myšlenku v podobě sofistikovanějších bayesovských sítí.

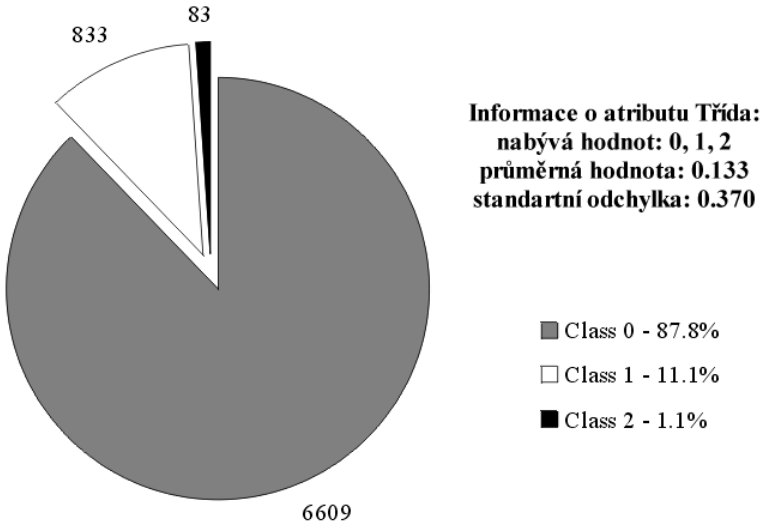
## Úloha predikce mortality - definice

Cílem úlohy predikce mortality je sestavit úspěšný prediktivní model pro předpověď výsledku kardiovaskulární intervence na základě pacientovy anamnézy a předoperačního stavu. Výsledek operace je explicitně definován na základě atributů NEARMISS+ a STATUS obsažených v registru MCZR. NEARMISS+ je počítaný paramter, který po provedení operace kvalitativně ohodnocuje její výsledek. Jedná se o binární parametr, hodnota 0 označuje úspěšnou intervenci, hodnota 1 intervenci neúspěšnou. Parametr STATUS označuje pacientův stav po operaci (1 – pacient je stále naživu, 2 – zemřel v průběhu operace, 3 – zemřel v průběhu hospitalizace, 4 – nutná reoperace, 5 – zemřel během 30 dnů po operaci, 6 – zemřel později než 30 dní po operaci, 7 – smrt způsobena jinými důvody než důvody vyvolanými operací). Pacienti byli klasifikováni do tří tříd dle následujícího klíče:

- Třída 0 - NEARMISS+ je 0, libovolná hodnota parametru STATUS,
- Třída 1 - NEARMISS+ je 1, STATUS je různý od 2 a 3,

- Třída 2 - NEARMISS+ je 1, STATUS je 2 nebo 3.

Úkolem je úspěšné sestavení takového klasifikátoru, který bude po svém natrénování schopen apriorně, tj. před započetím operace v době, kdy nejsou známy atributy NEARMISS+ ani STATUS, rozhodnout o přidělení pacienta do jedné z výše uvedených tříd.
Z důvodu značné odlišnosti údajů shromažďovaných o intervencích odlišných typů nelze sestavit jednotný prediktivní model pro všechny typy operací. V článku jsou prezentovány výsledky dvou prediktivních modelů budovaných nezávisle pro pacienty s intervencí typu bypass (značeno CABG) a pro pacienty, u kterých byla navíc provedena i náhrada chlopně (značeno RSVCE). Data typu RSVCE byla za období 1995-1999, data typu CABG pak za období 1995-1998.




**Obrázek 1 Rozdělení pacientů do tříd (typ dat CABG)**

Takto formulovaná úloha je úlohou klasifikační, kdy jednotlivé pacienty (záznamy o pacientech) přiřazujeme do disjunktních tříd. Úlohu je možné dále formulovat jako úlohu pravděpodobnostni (regresní), kdy každého z pacientů nepřiřadíme exkluzivně do jedné z předem definovaných tříd, ale přiřazujeme jim pravděpodobnosti příslušnosti do jednotlivých tříd. V těchto úlohách jsme z důvodu nerovnoměrného rozdělení pacientů do tříd a z důvodu přehlednosti vyhodnocení slučovali buď třídy 0 a 1 (pak se jedná o vlastní nemocniční mortalitu) nebo třídy 1 a 2 (v tomto případě se jedná o predikci atributu NEARMISS+).

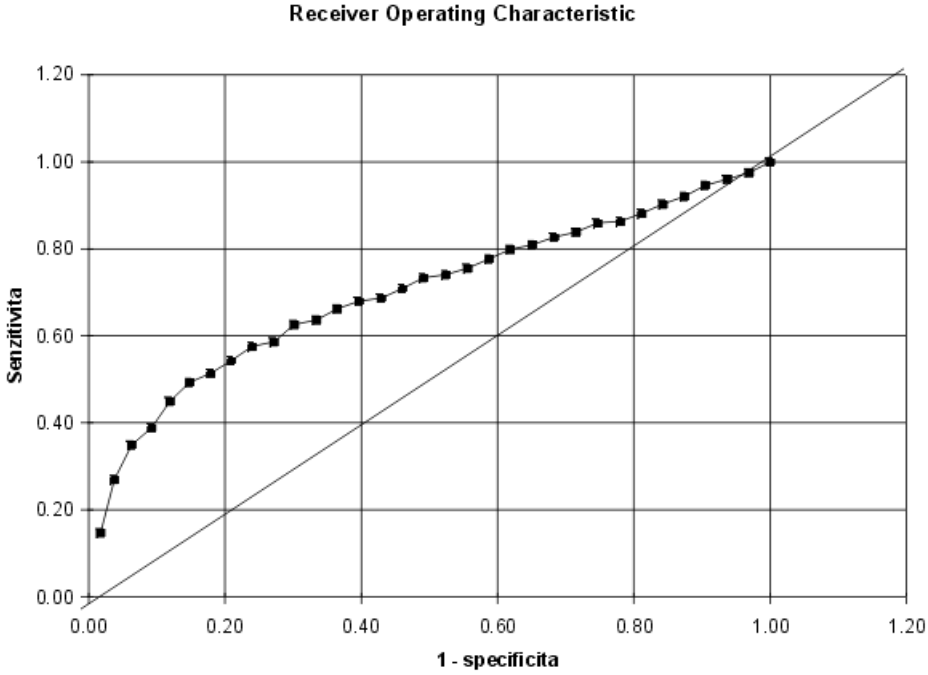
## Aktuální metodika řešení

Národní zdravotní registr kardiovaskulárních intervencí je spravován střediskem MEDICON [10]. Je provozován v prostředí expertního informačního systému PATS (The Patient Analysis and Tracking System) [11]. Tento systém nabízí kromě prostředků statistické analýzy také možnost budování prediktivních modelů, a to jednak krátkodobě orientovaných (slouží k předpovědi výsledku operace daného pacienta), ale i dlouhodobě orientovaných (se zaměřením na vývoj obecných jevů daného oboru). Pro konstrukci těchto modelů se užívá bayesovské analýzy. Výsledky bayesovského prediktoru integrovaného do systému PATS jsou v rámci prezentované studie považovány za referenční. Z pohledu terminologie učení se jedná o naivní Bayesův model, jeho prediktivní schopnosti jsou ale vylepšeny tím, že rizikové faktory, u nichž je známa vzájemná závislost, jsou sloučeny.

## Použité algoritmy a dosažené výsledky

První aplikovanou metodou byla technika tvorby rozhodovacích stromů. Byl použit algoritmus C5.0. Jedná se o vylepšenou modifikaci algoritmu C4.5 J.R.Quinlana [12]. Ve srovnání s předchozí verzí přináší dvě vylepšení, zvláště první z nich významně ovlivnilo způsob zpracování dat registru. C5.0 v první řadě umožňuje zavedení proměnlivé ceny ohodnocení chybné klasifikace příkladu. Často totiž požadujeme, aby některé chyby byly považovány za zvlášť nežádoucí, tj. ohodnoceny vyšším stupňem závažnosti při procesu učení. Kritériem úspěšnosti v průběhu procesu vytváření rozhodovacího stromu není samotný

počet chyb při klasifikaci, ale jejich souhrnná cena. Tímto způsobem lze potlačit význam jistého druhu chyb, v našem případě nízké úspěšnosti klasifikace minoritních tříd. Obecně totiž platí, že prediktivní modely mají tendenci konvergovat v případě nerovnoměrného rozdělení případů do tříd (viz Obrázek 1) k triviálnímu modelu, který by v našem případě predikoval konstantně třídu 0. Druhým vylepšením oproti C4.5 je implementace techniky zvané boosting navržené Freundem a Schapirem [3]. Jedná se o iterativní generování několika klasifikátorů, při kterém se daný klasifikátor zaměřuje na správnou klasifikaci příkladů chybně klasifikovaných svým předchůdcem.
Sesterským systémem k C5.0 je algoritmus generující regresní stromy Cubist [12]. Jedná se o modifikaci rozhodovacího stromu, kdy jsou v listech namísto třídy generovány regresní rovnice. V případě problematiky kardiochirurgických intervencí tyto rovnice predikují míru příslušnosti k jednotlivým třídám. V případě regresních stromů snižuje kvalitu učení fakt, že u trénovacích příkladů je příslušnost k jednotlivým třídám vždy jednoznačná a dochází tedy ke změně formátu výstupu.
Jako neurální prediktor byl použit vícevrstvý perceptron (multi-layer perceptron MLP), konkrétně algoritmus, který je součástí nástroje DataEngine [13].
iBARET (Instance-Based REasoning Tool) reprezentuje implementaci systému případového usuzování. Jedná se o univerzální prediktivní systém použitelný pro úlohy, ve kterých jsou jednotlivé příklady popsané vektorem numerických nebo symbolických hodnot. Systém byl vytvořen na katedře kybernetiky, FEL ČVUT [6], výsledky jeho aplikace na data střediska MEDICON jsou podrobněji popsány ve [4] a [5].
Ve všech případech kromě bayesovského modelu byla zajištěna nezávislost trénovacích a testovacích dat, čímž bylo zabráněno přeučení. Nebezpečí přeučení je největší zejména v případě neuronových sítí a rozhodovacích stromů u Bayesova modelu by odhad chyby neměl být výrazně zkreslen.
Kvalita jednotlivých prediktivních modelů byla hodnocena prostřednictvím ROC (Receiver Operating Characteristic) křivek [2]. Jedná se o metodu mající původ v radiologii, často využívanou i ve zdravotnictví. V případě posuzování rozhodovacích metod je definována jako křivka vyjadřující vztah mezi počtem správně pozitivně a nesprávně pozitivně ohodnocených pacientů na různých úrovních prahu oddělujícího pozitivní a negativní klasifikace. Za pozitivní pacienty považujeme pacienty patřící do tříd 1 a 2 (případně pouze 2), za negativní pak všechny ostatní pacienty. Kvalita prediktivního modelu odpovídá ploše pod ROC křivkou. Příklad této křivky je uveden na Obrázek 2.




**Obrázek 2 Příklad ROC křivky**

Na příkladě rozhodovacího stromu lze nejsnáze upozornit na teoretická úskalí při vytváření prediktivních modelů. Přestože registr obsahuje tisíce záznamů pro oba sledované typy intervencí, každý ze záznamů se skládá z několika deítek atributů (přibližně 40). Některé z těchto atributů mohou navíc teoreticky nabývat

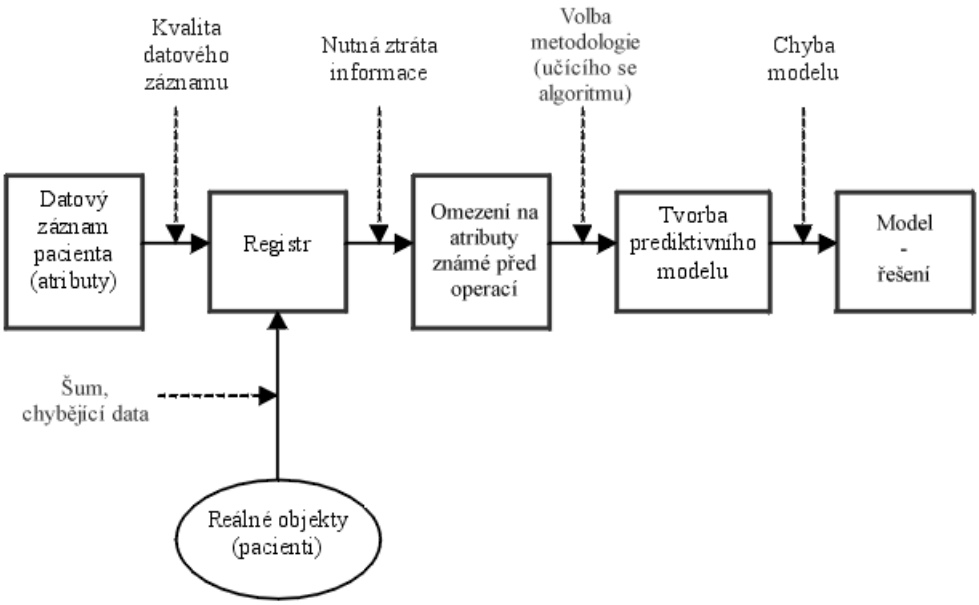
široké škály hodnot a pro účely učení musí být transformovány na více jednodušších, nejlépe dichotomických atributů. Zároveň platí tzv. PAC teorém (probably approximately correct learning theorem) [8], který definuje minimální počet trénovacích příkladů pro danou hloubku rozhodovacího stromu, požadovanou maximální chybu učení a složitost prohledávaného prostoru hypotéz. Složitost prostoru hypotéz je dána právě počtem atributů. Dle PAC teorému je pro 50 atributů, maximální chybu 10% a hloubku stromu 4 doporučený počet příkladů 63 miliónů. PAC teorém navíc předpokládá rovnoměrné dělení příkladů do jednotlivých tříd. Je tedy evidentní, že počet atributů musí být výrazně redukován.
Přehled výsledků dosažených aplikací jednotlivých metod je uveden v Tabulka 1. Bayesův model je referenční, u algoritmu C5.0 je uveden bodový odhad velikost ROC plochy (jedná se o klasifikační algoritmus, nelze volit různé hodnoty prahu oddělujícího pozitivní a negativní příklady).

| Algoritmus \ Typ dat | RSVCE (6.671 záznamů) | CABG (7.525 záznamů) |
|---|---|---|
| Bayes* | 0.709 | 0.711 |
| MLP | 0.715 | 0.720 |
| iBARET | 0.696 | 0.702 |
| Cubist | 0.652 | 0.672 |
| C5.0** | 0.621 | 0.612 |

**Tabulka 1 Přehled dosažených výsledků – ROC plochy**

## Zhodnocení výsledků

V rámci hodnocení prediktorů pomocí velikosti ROC plochy můžeme daný prediktor považovat za ideální pokud je velikost jeho ROC plochy 1, nulová plocha indikuje ideální invertovaný prediktor. Naopak náhodný prediktor, jehož výsledek není nijak korelován s reálnou příslušností pacienta ke třídě, vykazuje ROC plochu o velikosti 0.5. Z tabulky výsledků je zřejmé, že jednotlivé modely vykazují významnou korelaci predikcí se skutečnými výsledky intervencí. Otázkou je však stanovení hranice, od které lze prediktivní model považovat za spolehlivý. Dle praktických zkušeností lze za vágní hranici spolehlivosti považovat plochu o velikosti 0.8.
Ani jeden z prediktivních modelů této hranice nedosahuje. Obrázek 3 shrnuje možné důvody. Ke ztrátě informace dochází už samotnou definicí datového záznamu pacienta. Vlivem šumu při měření, popřípadě vlivem nesprávně vyplněných či chybějících údajů dochází k dalším informačním ztrátám. Významným omezením je také možnost použití pouze informací známých před operací, nikoli v průběhu operace či po ní. Konečně další chybu přináší samotný model, případně jeho nevhodná volba. Dalším důvodem může být i relativní nedostatek dat. Registr sice obsahuje tisíce pacientů pro oba sledované typy intervencí, podíl pozitivních příkladů (třída 2 resp. 1) je však velmi malý.
Bylo testováno 5 principiálně významně odlišných metod tvorby prediktivních modelů. Fakt, že ani jedna z nich nedosáhla výše zmiňovaného prahu spolehlivosti naznačuje, že spolehlivá preoperativní predikce mortality je nad stávajícím obsahem registrů obtížně realizovatelná.




**Obrázek 3 Důvody ztráty přesnosti při konstrukci prediktivního modelu**

## Závěr

Navzdory tomu, že zdravotní registry shromažďují velké množství záznamů (desetitisíce až statisíce pro jednotlivé typy operací) a tyto informace mají značné možnosti využití, některé typy úloh přesto nemohou být na jejich základě spolehlivě řešeny. Prezentované výsledky naznačují, že do výše zmíněné třídy úloh patří také predikce mortality v rámci kardiovaskulárních intervencí. Přestože většina použitých metod strojového učení generuje predikce významně korelující se skutečnými výsledky operací, ani jedna z aplikovaných metod negeneruje predikce takové kvality, aby byly prediktivní modely využitelné např. jako konzultační expertní systémy. Nejlepšího výsledku bylo dosaženo pomocí neuronových sítí, rozdíl oproti Bayesovu prediktivnímu modelu však není statisticky signifikantní.
Na druhou stranu vytvořené modely z hlediska extrakce znalostí významně napomáhají v identifikaci a stratifikaci jednotlivých rizikových faktorů, stejně jako v kvantifikaci jejich případných vzájemných závislostí.

Článek vznikl za podpory výzkumného záměru Transdisciplinární výzkum v oblasti biomedicínského inženýrství MSM 210000012.

## Literatura

[1] Weiss, S., M., Indurkhya, N.: Predictive Data Mining (a practical guide), Morgan Kaufmann Publishers, 1998.

[2] Hanley, J., A., McNeill, B., J.: The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, p.29-36, 1982.

[3] Freund, Y., Schapire, R.,E.: Experiments with a New Boosting Algorithm, in Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148-156, Morgan Kaufmann Publishers, 1996.

[4] Klema, J., Lhotská, L., Palouš J., Štěpánková, O.: Instance-Based Modeling in Medical Systems. Cybernetics and Systems 2000, Proceedings of 15th European Meeting on Cybernetics and Systems Research (EMCSR), (R. Trappl, ed.), Vienna, 2000, Austrian Society for Cybernetic Studies, ISBN 3 85206 151 2, pp. 365-370, 2000.

[5] Klema, J., Štěpánková, O., Mikšovský, P.: Predictive Model of Heart Operation Result – Built on Merged National Registry on Cardiovascular Interventions of MEDICON Center. Technical

Report 72/98 of The Gerstner laboratory CTU Prague & FAW Linz – Hagenberg – Praha- Wien, 1998.

[6] Palouš, J.: Využití případového usuzování pro lékařské aplikace. Diplomová práce, FEL ČVUT, Praha, 2000.

[7] Mařík, V., Štěpánková, O., Lažanský, J.: Umělá inteligence I., Academia, 1993.

[8] Mařík, V., Štěpánková, O., Lažanský, J.: Umělá inteligence III., Academia, 2001.

[9] Štěpánková, O., Hetmerová, A., Kraus, J.: Některé možnosti použití strojového učení v lékařství, Lékař a technika 29/1998, pp. 56-62.

[10] www stránky střediska MEDICON, www.medicon.cz.

[11] www stránky společnosti Axis Clinical Software, Inc., www.axisclinical.com

[12] www stránky společnosti RuleQuest Research, www.rulequest.com.

[13] www stránky společnosti Management Intelligenter Technologien, GmbH pro produkt DataEngine, www.dataengine.de.

Ing. Jiří Klema

Gerstnerova laboratoř pro inteligentní rozhodování a řízení

katedra kybernetiky, FEL ČVUT

Technická 2

166 27, Praha 6

e-mail: klema@labe.felk.cvut.cz

tel.: 2435 7392