

## WEKA – soubor algoritmů strojového učení



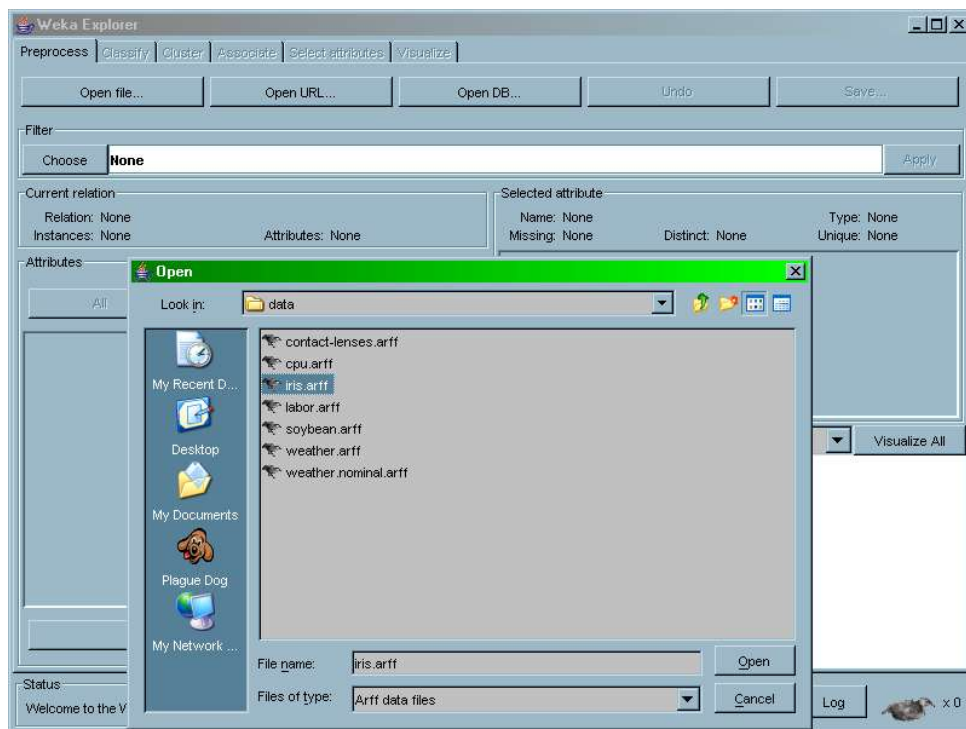
**WEKA**, neboli *Waikato Environment for Knowledge Analysis*, je poměrně rozsáhlý soubor algoritmů strojového učení vyvinutý na the University of Waikato in New Zealand (*weka* je zároveň jméno unikátního nelétajícího ptáka, který existuje pouze na Novém Zélandě). Systém je neustále rozšiřován, laděn a doplňován. WEKA je implementován v jazyce Java a v současné době disponuje rovněž pohodlným uživatelským grafickým rozhraním (GUI). Pro cvičení v předmětu *Strojové učení* je systém WEKA nainstalován v příslušných počítačových učebnách jak pod Windows, tak i pod Linuxem.

Program WEKA verze 3.4 (podzim 2004) je rovněž k dispozici na URL <http://www.cs.waikato.ac.nz/ml/weka> odkud si jej lze stáhnout včetně Java-2 1.4.2 (případně vyšší verze) pro různé typy operačních systémů podporujících jazyk Java. Na uvedené URL jsou také podrobnější informace (pro případné zájemce). S uvedeným software úzce souvisí také kniha od dvou autorů: Ian H. Witten a Eibe Frank: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2000, Morgan Kaufmann Publishers. Součástí software je soubor v pdf, který obsahuje podstatnou kapitolu č. 8 (je k dispozici rovněž pro cvičení s WEKA). Vzhledem k neustále pokračujícímu rozvoji je nutno vzít na vědomí, že WEKA obsahuje také nějaké chyby, které jsou postupně v dalších verzích systému odstraňovány, takže zájemcům o aplikace strojového učení na řešení nejrozumnějších problémů se doporučuje občasná návštěva na výše uvedené URL.

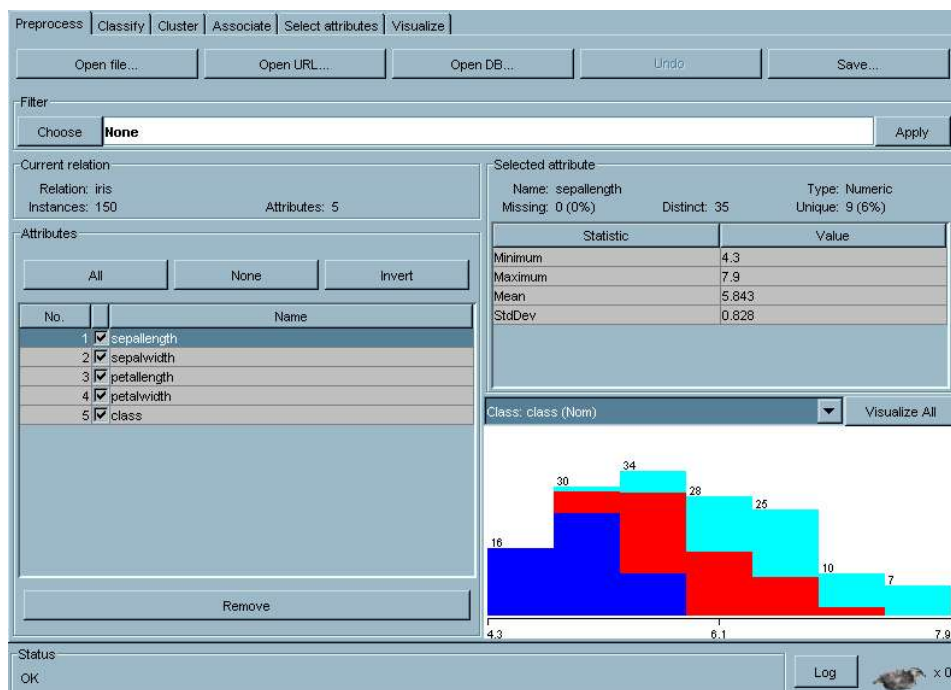
Cílem cvičení zaměřených na systém WEKA je jednak se seznámit se základním použitím programu a s jeho vlastnostmi, a dále vyzkoušet si různé algoritmy na různých datech, včetně např. porovnání výsledků různých algoritmů aplikovaných na tatáž data, apod. Hlubší zájemci mohou získat velmi fundované a rozsáhlé vědomosti pomocí experimentování; systém WEKA je zaměřen na praktickou stránku strojového učení, protože obdobný sjednocený a široký soubor různých algoritmů dané oblasti jinde k dispozici není.

Cvičení s WEKA vyžaduje po studentech a studentkách, aby se seznámili s popisem systému, jehož používání je však velmi jednoduché navzdory značnému množství funkcí a možností. Součástí instalovaného software by (kromě osmé kapitoly *Tutorial.pdf* výše zmíněné knihy) měl být také soubor *ExplorerGuide.pdf* s detailnějším popisem způsobu použití WEKA (oba pdf soubory jsou v angličtině a jsou přiloženy k zadání cvičení). Kromě toho je k dispozici český text, který vznikl jako součást bakalářské práce studenta FI MU pana Bc. Jakuba Širokého na jaře 2004. V tomto textu je rovněž detailní popis použití systému z hlediska vybraných algoritmů a dat, včetně některých užitečných doporučení. Bc. práce je k zadání cvičení také přiložena jako soubor *WEKA\_Bc\_Jakub\_Siroky.pdf*.

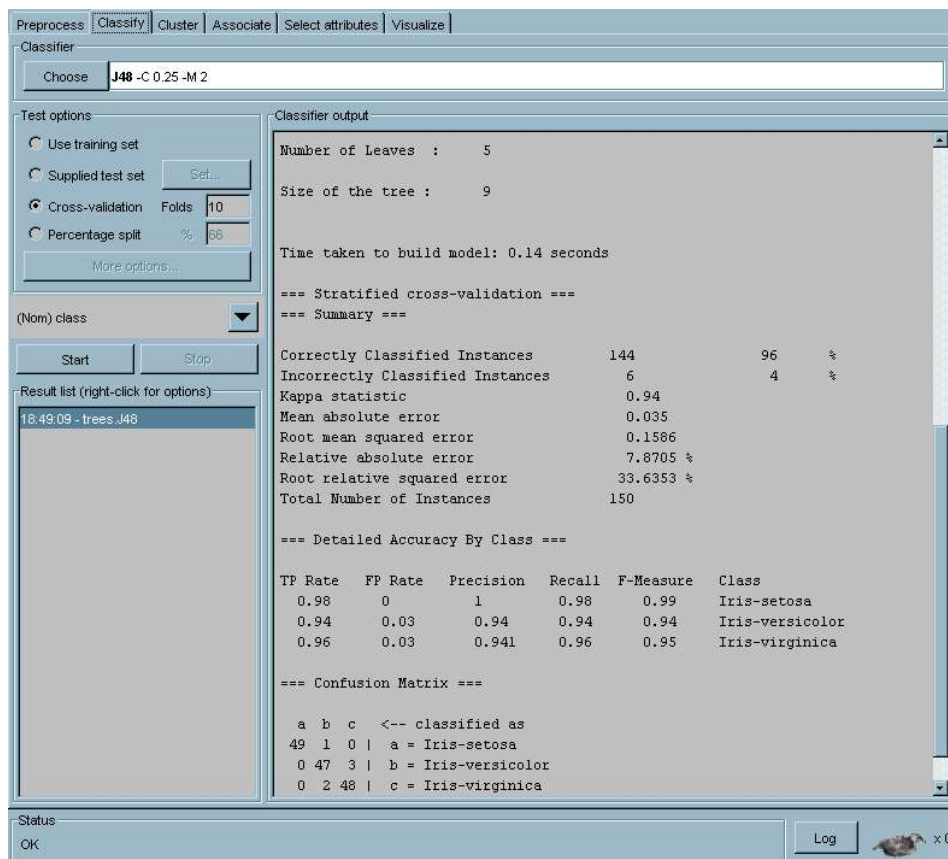
K základnímu seznámení se s WEKA je zapotřebí program spustit a na obrazovce se objeví úvodní obrázek uvedený na začátku tohoto textu. Doporučuje se začít s částí systému *Explorer*, kde je k dispozici mnoho funkcí (zbývající tři možnosti jsou také popsány v citované Bc. práci a zájemci si je mohou vyzkoušet). Trénovací datový soubor se zadá pomocí menu *Preprocess* (předzpracování) a *Open file* – WEKA má jednak vlastní (jednoduchý) textový datový formát *\*.arff*, jednak umí zpracovat i jiné formáty (např. studentům již známý formát *\*.names* používaný pro *See5/c5*, nebo *\*.csv*). Formát *arff* je posán v přiložených PDF zdrojích a lze jej prozkoumat i v datových souborech. Např. pro data *Iris* použitá pro *See5/c5* v prvním cvičení může dialog vypadat takto:



V dalším kroku je nutno zadat, které atributy budou použity, např. všechny pomocí *All*. *Explorer* zobrazí mj. i statistické informace o zadaných datech, což může být velmi užitečné při hledání vhodného postupu zpracování dat (např. výběr algoritmu, atd.):



Dále je možné nastavit řadu různých parametrů – viz příloženou dokumentaci. Dané cvičení směřuje především k úvodu do používání WEKA, a úvodní příklad s daty *Iris* lze spustit přechodem na menu *Classify* (vedle *Preprocess*), kde v části *Classifier* (klasifikátor) pomocí *Choose* (výběr) je nutno zvolit učící algoritmus. Ve WEKA je obdobou *See5/c5* algoritmus *J48* (vyvinutý z původního algoritmu *c4.5*, tj. generování rozhodovacích stromů pomocí minimalizace entropie – WEKA nabízí rovněž ještě dalšího předchůdce, *ID3*, a další). Pesimistický odhad budoucí chyby klasifikátoru lze zjistit např. pomocí parametru *Cross-validation*, a podobně lze zadat případné další parametry algoritmu (pro různé algoritmy jsou k dispozici samozřejmě různé parametry). Např. pro desetinásobné křížové ověřování získá uživatel (po zahájení trénování tlačítkem *Start*) následující výsledek (pozn.: *Classifier output* je na ilustraci zobrazen jen částečně, ke konci):



Studenti a studentky si vyzkoušejí používání WEKA pro různá data (některé algoritmy ovšem vyžadují pouze omezený typ dat, např. ID3 na rozdíl od J48 umí pracovat pouze se symbolickými daty, nikoliv numerickými spojitými, ale program na takové a podobné nedostatky v případě výskytu upozorní).

Dále pro tatož data se vyzkoušejí různé algoritmy, např. *Naivní Bayes*, nebo ze skupiny *lazy* IB1 a IBk (tj. 1-NN a k-NN nejbližší soused; parametry algoritmu se zadávají přes příslušné dialogové okno vyvolané cvaknutím myši do řádku vedle tlačítka *Choose*, kde je zobrazen název příslušné Java třídy s parametry, tj. zvolený algoritmus). Pozn.: *lazy* (líné) algoritmy jsou ty, které při trénování vlastně nic nedělají, pouze ukládají příklady do paměti, a pilné (indukují) jsou až v okamžiku použití pro klasifikaci/aproximaci.

Doporučit pro začátek lze např. data *letter* (příklady písmen A, B, ..., Z); další data (a na ně aplikované algoritmy) si je možno rovněž vyzkoušet.

Je vhodné např. zvolit si určitá data a parametry experimentů (*cross-validation*, ...), spustit na trénovacích datech různé algoritmy a výsledky porovnat. Takto lze mj. zjistit, který učící algoritmus v pozdějším běžném použití je schopen dávat co nejlepší očekávané výsledky na základě trénovacích dat.

WEKA obsahuje řadu užitečných funkcí, například *Visualize*.