přičemž S_x a S_y se spočítají analogicky jako S_p a S_s uvedené v odstavci 6.1.10. V našem případě budou oněmi sadami čísel správnosti (resp. chyby) stanovené například násobnou křížovou validací pro dva různé modely. Model A bude lepší než model B, pokud t-testem zjistíme, že jeho průměrná správnost je statisticky významně vyšší, tedy že

$$t(Acc_A, Acc_B) \geq t(1 - \alpha/2, m + n - 2),$$

kde $t(1 - \alpha/2, m + n - 2)$ je $(1 - \alpha/2)$-kvantil Studentova t rozložení s $m+n-2$ stupni volnosti (Havránek, 1993).

6.3.2 Použití křivek ROC

Křivky ROC umožňují zachytit chování modelu (klasifikátoru) bez ohledu na rozdělení tříd a cen. Očekávaná cena klasifikace modelu, který odpovídá bodu (TP, FP) křivky, je

$$Cost = p(p) \,.\, FP[\%] \,.\, c(P,n) + p(n) \,.\, FN[\%] \,.\, c(N,p),$$

kde $p(p)$ a $p(n)$ jsou apriorní pravděpodobnosti (relativní četnosti) pozitivních a negativních příkladů, $c(P,n)$ a $c(N,p)$ udávají ceny za chybnou klasifikaci a $FP[\%]$ a $FN[\%]$ značí poměry správně pozitivních a správně negativních klasifikací. Dva body (TP_1, FP_1), (TP_2, FP_2) budou odpovídat stejně kvalitním modelům, pokud

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{p(n)c(P,n)}{p(p)c(N,p)}.$$

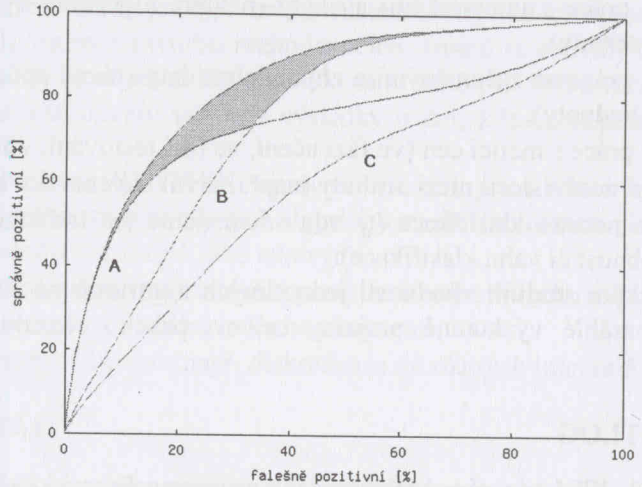
Tato rovnice definuje směrnici linie stejného výkonu (iso-performance lines); všechny modely ležící na jedné linii mají stejné očekávané ceny klasifikace.

Obrázek 138 ukazuje křivky ROC pro tři modely. Je zřejmé, že model C je nejhorší, protože jeho křivka leží všude „pod" křivkami zbývajících modelů. Jiné je to pro modely A a B. Při některých strategiích rozhodování bude vhodnější model A (minimalizujeme $FP[\%]$ i za cenu nižší hodnoty $TP[\%]$), při jiných strategiích bude vhodnější model B (maximalizujeme $TP[\%]$ i za cenu vyšší hodnoty $FP[\%]$). Pro hodnoty směrnic linií stejného výkonu v intervalu $\langle 0,1\rangle$ (tedy např. směrnice 0,1 pro $p(n)/p(p) = 1$ a $c(P,n)/c(N,p) = 0,1$) bude lepší model B, pro hodnoty směrnic v intervalu $\langle 1,\infty)$ (např. směrnice 2 pro $p(n)/p(p) = 10$ a $c(P,n)/c(N,p) = 0,2$) bude lepší model A.

Jak tedy nalézt metodu, která bude optimální pro danou strategii (pro danou matici cen, resp. poměr tříd)? K tomu se použije tzv. *konvexní obal* (convex hull) prostoru ROC, který „obepíná" dané křivky ROC. Klasifikátor, který bude ležet na tomto obalu, bude (pro danou strategii) optimální. Lze totiž ukázat, že

neleží-li bod křivky ROC na konvexním obalu, lze pro libovolnou rodinu linií stejného výkonu (tedy pro linie se stejnou směrnicí) nalézt bod, který leží na linii se stejnou směrnicí, ale protíná osu $TP[\%]$ ve větší hodnotě.

Vidíme tedy, že pro určité strategie nebude ani model A, ani model B optimální, protože příslušné křivky ROC leží „pod" konvexním obalem (tato část je v grafu vyznačena šedou barvou). Zlepšení klasifikace lze pak dosáhnout k o m b i n a c í obou modelů.



Obr. 138 Křivky ROC s konvexním obalem.

6.3.3 Occamova břitva

V případě shody správnosti (chyby) vstupují do hry další kritéria. Nejznámějším je tzv. *Occamova břitva.* Jde o filozofický předpoklad, který říká, že nejlepší vědecká teorie je ta nejjednodušší, která popisuje všechna fakta. Převedeno na porovnávání modelů jde o kritérium, které říká, že lepší je menší model (méně pravidel, menší strom apod.). V teorii učících se systémů má tento předpoklad jednoduchosti podobu principu minimální délky popisu (minimum description length, MDL), o kterém byla již zmínka v souvislosti s bayesovským klasifikátorem.

6.4 Volba nejvhodnějšího algoritmu

Vzhledem k tomu, že neexistuje algoritmus, který by předčil ostatní na libovolných datech[201], dostává se do popředí otázka jak dopředu poznat, který

[201] Tato skutečnost je známa pod názvem no free lunch teorem.