August 18, 2015

Pois é, você deve estar enviesando seu problema. Dependendo das características, é capaz que seja possível corrigir esse viés. Vou tentar formalizar um pouco aqui.

Digamos que sua base original é

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

e sua base obtida online é

$$(X'_1, Y'_1), \ldots, (X'_{n'}, Y'_{n'}),$$

onde X representam suas covariáveis e Y a resposta.

A base estar viesada essencialmente significa que a distribuição de (X, Y) é diferente da distribuição de (X', Y'). O que isso implica? Bem, sem suposições adicionais, isso implica que é muito difícil usar a segunda base, pois ela pode ter características completamente distintas. Portanto vamos ter que fazer alguma suposição extra. Talvez uma suposição razoável para o seu problema (talvez!!!) seja a chamada *covariate shift*. Essencialmente, assumimos que

$$Y|X = x \sim Y'|X' = x,$$

isto é, fixado o valor das covariáveis, a distribuição da variável resposta é a mesma em ambos os conjuntos. Ou seja, permitimos que a distribuição marginal das covariáveis seja diferente (ex: se é muito mais fácil fazer classificações corretas para $x = (1, 1, 4)$, então haverá muito mais amostras com valores próximos a esse no segundo banco).

O seu problema tem algo interessante que eu nunca havia visto antes: mesmo que essa suposição seja falsa, ela pode levar a classificadores melhores se bem usada, e usando o conjunto de testes vamos saber isso. Então vou sugerir algo aqui (se você tiver tempo e interesse) que pode melhorar esses classificadores.

Primeiro, vamos formalizar mais um pouco. Vou assumir por simplicidade que suas variáveis são contínuas. Vamos chamar de $f(x, y)$ a distribuição no banco original (distribuição de interesse) e $g(x, y)$ a distribuição no banco novo. A suposição covariate shift assume que $f(y|x) = g(y|x)$, mas que as marginais podem ser diferentes. Qual seu interesse do ponto de vista estatístico? Encontrar

uma função de classificação $h(x)$ que tenha *risco baixo*, i.e., tal que

$$\mathbb{E}[L(h(X), Y)]$$

seja pequeno. Aqui, $L(h(X), Y)$ é a perda que você está usando, por exemplo $L(h(X), Y) = \mathbb{I}(h(X) \neq Y)$. Esta esperança é tomada com relação à distribuição de interesse, isto é,

$$\mathbb{E}_f[L(h(X), Y)] = \int L(h(x), y) f(x, y) dx dy.$$

Quase todos os métodos de aprendizado de máquina tentam minimizar essa quantidade. É isso que fazemos quando escolheremos tuning parameters por validação cruzada! O primeiro problema no seu caso é que a segunda amostra tem uma distribuição diferente, logo o que é bom para a segunda amostra não necessariamente é bom para a primeira. Imaginando que a segunda base é muito maior que a primeira, se consideramos todas as amostras conjuntamente de forma ingênua, encontraremos h que tem $\mathbb{E}_g[L(h(X), Y)]$ baixo, mas não necessariamente $\mathbb{E}_f[L(h(X), Y)]$. Mas podemos corrigir isso usando nossa suposição:

$$\begin{aligned}
\mathbb{E}_f[L(h(X), Y)] &= \int L(h(x), y) f(x, y) dx dy = \int L(h(x), y) f(y|x) f(x) dx dy = \\
&= \int L(h(x), y) g(y|x) \frac{f(x)}{g(x)} g(x) dx dy = \int L(h(x), y) g(x, y) \beta(x) dx dy \\
&= \mathbb{E}_g[L(h(X), Y)\beta(X)]
\end{aligned}$$

com $\beta(x) = \frac{f(x)}{g(x)}$. Ou seja, estimando $\beta(x)$, podemos corrigir o risco estimado no segundo grupo para conseguir classificadores que sejam bons no primeiro grupo. No caso específico de random forests, acho que você fazer isso incluindo um peso para cada amostra proporcional a $\beta(x)$ se a amostra pertence ao segundo banco e um caso contrário. Lembro que árvores do pacote tree tem um parâmetro indicando o peso de cada observação, acredito que florestas também devam ter. Note a intuição desses pesos: se x tem "probabilidade" alta no conjunto 2 mas probabilidade baixa no conjunto 1 (distribuição de interesse), o peso que ele recebe é pequeno ($f(x)$ é pequeno e $g(x)$ é grande). Faz sentido! Essa é uma região do espaço amostral que dificilmente irá aparecer em novos exemplo; o fato de você ter vários desses em seu banco é simplesmente porque é uma região de fácil classificação.

Existem vários métodos para estimar $\beta(x)$. Coincidentemente ou não trabalhei com isso recentemente, mas em um contexto diferente. Veja equação 3 de https://www.dropbox.com/s/lowci637oak1cnc/main.pdf?dl=0 para um exemplo de estimador. Dependendo do seu problema, outros estimadores tem chance maior de funcionar melhor. Se você estiver interessado nessa direção, me fale que entro mais em detalhes, mas é bem simples.

Por fim: talvez valha a pena você testar métodos de aprendizado semi-supervisionado, aí você poderia usar o banco novo todo (sem contudo usar as classificações, mas ele não seria viesado).