Aprendizado de Semi-Supervisionado de Classificadores Bayesianos Utilizando Testes de Independência

Marcelo C. Cirelo[1], **Fabio G. Cozman**[1]

[1] Escola Politécnica da Universidade de São Paulo
Av. Prof. Luciano Gualberto, 158, tv. 3 – 05508-900 São Paulo, SP

marcelo.cirelo@poli.usp.br, fgcozman@usp.br

Abstract. *In this work we present algorithms for learning Bayesian networks from labeled and unlabeled data. We compare semi-supervised versions of standard Bayesian classifiers like Naive Bayes and TAN to a classifier based on the CBL1 algorithm [Cheng et al., 1997]. Our contribution is the development of a EM-like framework for the CBL1 algorithm. There is empirical evidence that the presented classifier performs better than TAN and Naive Bayes when dealing with labeled and unlabeled data.*

Resumo. *Neste trabalho são apresentados métodos para o aprendizado de classificadores Bayesianos a partir de bases que contenham dados classificados e não-classificados. Foram comparadas versões modificadas para o aprendizado semi-supervisionado dos algoritmos Naive Bayes e TAN com um classificador baseado no algoritmo CBL1 [Cheng et al., 1997]. Nossa contribuição é o desenvolvimento de um método para utilizar o CBL1 com EM. Os resultados empíricos mostram que o algoritmo proposto tem desempenho superior aos algoritmos Naive Bayes e TAN para o aprendizado a partir de dados classificados e não-classificados.*

1. Introdução

Um ponto crítico do projeto de classificadores é a obtenção de uma quantidade razoável de dados classificados para treino, seja devido ao seu custo ou a outra dificuldade de ordem prática. Esse problema poderia ser minimizado aumentando-se a base de dados existente com dados não classificados, que são abundantes em aplicações como, por exemplo, categorização de páginas da internet.

[Baluja, 1999, Nigam et al., 2000] apresentam casos em que classificadores Bayesianos treinados com dados classificados e não classificados têm desempenho superior ao que seria obtido caso os dados não classificados tivessem sido simplesmente ignorados. Esses trabalhos, no entanto, assumem modelos fixos para os dados, que se aplicam bem aos problemas apresentados, mas não necessariamente serão úteis para bases de dados provenientes de domínios diversos.

Neste artigo apresentamos um método que, utilizando um algoritmo de aprendizado de estruturas baseado em testes de independência, escolhe o modelo a partir dos dados classificados e dos dados não-classificados disponíveis. Foram realizados testes

com bases provenientes do repositório da UCI; os resultados foram comparados com os classificadores Bayesianos mais comuns na literatura: o Naive Bayes e o TAN.

O texto é organizado da seguinte forma: na Seção 2, são descritos os principais algoritmos para o aprendizado supervisionado de classificadores. A Seção 3 descreve o aprendizado semi-supervisionado e detalha o algoritmo proposto. A Seção 4 contém a descrição dos experimentos realizados e a discussão dos resultados. Finalmente, a Seção 5 contém conclusões, considerações finais e trabalhos futuros.

2. Aprendizado Supervisionado de Redes Bayesianas para a Classificação

Esta seção apresenta o aprendizado de classificadores a partir de bases de dados completamente classificadas.

A tarefa de classificação consiste em associar um rótulo c ao vetor de observações $\mathbf{x}$. Uma abordagem probabilística envolve a modelagem da distribuição conjunta do espaço de observações e da classe. Conhecida a distribuição $P(C,\mathbf{X})$, a escolha da classe que maximiza a taxa de acerto esperada do classificador é feita utilizando-se a regra de decisão de Bayes [Duda e Hart, 1973]. $P(C,\mathbf{X})$ em geral não é conhecida e pode ser estimada a partir de uma base de dados D. Neste texto, letras maiúsculas são utilizadas para representar variáveis aleatórias discretas, letras minúsculas indicam variáveis observadas, $|\,.\,|$ indica o número de valores possíveis de dada variável aleatória pode assumir.

Redes Bayesianas, que constituem a base dos algoritmos apresentados neste artigo, permitem a representação de distribuições conjuntas de probabilidades e podem ser representadas pelo par $\langle G,\theta\rangle$, onde G é um grafo acíclico direcionado (DAG, na sigla em inglês) que contém as relações entre as variáveis (nós do grafo) e θ o conjunto de parâmetros que quantifica a distribuição. Neste texto, o termo "estrutura" e "modelo" da rede Bayesiana refere-se à topologia do grafo G. Denotamos por $Pa(X)$ o conjunto de variáveis de onde se originam os arcos que terminam em X, esse conjunto constitui as variáveis *pais* de X.

Os classificadores Bayesianos generativos mais simples[1], conhecidos como Naive Bayes (NB), partem da hipótese que todos os atributos são independentes dado a classe. Apesar dessa premissa pouco realista, classificadores NB são eficazes em muitos casos práticos de aprendizado supervisionado [Domingos e Pazzani, 1997].

A hipótese de independência do Naive Bayes pode ser suavizada através da adição de arcos que representam relações de dependência entre os atributos dado a classe. É sabido que a recuperação da melhor estrutura, restrita ao espaço de estruturas do tipo árvores, pode ser feita em tempo polinomial [Chow e Liu, 1968]. O algoritmo *Tree Augmented Naive Bayes* [Friedman et al., 1997] (TAN) utiliza essa propriedade na construção de classificadores.

Como muitas relações de dependência entre as variáveis não podem ser representadas nem mesmo por estruturas tipo TAN, é necessária a construção de modelos mais complexos que permitam que cada nó da rede (exceto a classe) tenha um número arbitrário de pais. Como o aprendizado de estruturas arbitrárias pode ser bastante custoso computacionalmente, utiliza-se algum conhecimento *a priori*, como a ordenação prévia

[1]Nas estruturas generativas a classe é necessariamente um nó raiz.

das variáveis, para possível tornar o aprendizado exeqüível. Essa é a abordagem adotada em [Cheng e Greiner, 1999] para o aprendizado de estruturas para classificação, esses autores utilizam uma versão do algoritmo CBL1 para gerar classificadores que, nos exemplos apresentados, superam NB e TAN.

O algoritmo CBL1 é divido em três fases: na primeira, cria-se um esboço da rede utilizando-se testes de independência, de maneira semelhante ao descrito no artigo de Chow e Liu; na segunda fase realiza-se testes de independência condicional para decidir quais arcos devem ser adicionados; finalmente, como as duas fases anteriores podem adicionar arcos desnecessários, na terceira fase, são realizados novos testes de independência para a remoção dos arcos sobressalentes.

O CBL1 foi utilizado para a tarefa de classificação em [Cheng e Greiner, 1999]. Neste trabalho adotamos uma abordagem um pouco diferente: a primeira fase do algoritmo é substituída por uma estrutura simples como o NB, e, na terceira fase, pode-se remover os arcos que ligam a classe aos atributos.

Outros algoritmos de aprendizado de Redes Bayesianas buscam maximizar *scores* como a verossimilhança ou *Bayesian score*, porém resultados anteriores indicam que esse procedimento, apesar de produzir redes de alto *score*, não garantem bons classificadores. Os textos [Friedman et al., 1997, Cowell, 2001] apresentam uma discussão sobre esse ponto. Outros trabalhos focam em *scores* específicos para classificação como o algoritmo SSS [Cohen et al., 2003] que realiza uma busca guiada pela taxa de acerto do classificador.

3. Aprendizado Semi-Supervisionado

No aprendizado semi-supervisionado dispomos de uma base de dados classificados D' que é complementada com dados não-classificados D''. Da mesma forma que no caso de aprendizado supervisionado, buscamos encontrar o conjunto de parâmetros θ que maximize a verossimilhança de rede com relação aos dados, considerando que o modelo da distribuição seja conhecido (ou assumido) *a priori*.

Dados não-classificados trazem informação apenas sobre a distribuição marginal dos atributos $P(\mathbf{X})$. Em estruturas generativas $P(\mathbf{X})$ pode ser decomposta em $P(\mathbf{X}) = \sum_{i=1}^{|C|} P(C) \prod_{\mathbf{X}} P(\mathbf{X} \mid Pa(\mathbf{X}))$. Note que, para que seja possível tirar proveito dos dados não-classificados, deve-se restringir a família de estruturas à classe de estruturas generativas. Redes de diagnóstico (em que os atributos podem ser pais da classe) são amplamente utilizadas em aprendizado supervisionado de classificadores. Para essas estruturas apenas a distribuição $P(C \mid \mathbf{X})$ é parametrizada, descartando a distribuição marginal dos atributos $P(\mathbf{X})$.

A função de verossimilhança de uma base mista que contenha tanto exemplos completos quanto exemplos não-classificados é dada pela equação 1.

$$\sum_{i=1}^{|D'|} log\left(P(C=y_i)\prod_{\mathbf{x}_i} P(x \mid Pa(x_i),\theta)\right) + \sum_{i=1}^{|D''|} log\left(\sum_{j=1}^{|C|} P(c_{ij} \mid \theta)\prod_{\mathbf{x}_i} P(x \mid Pa(x_i),\theta)\right), \tag{1}$$

onde y_i indica a classe correta do exemplo.

Tabela 1: O algoritmo EM-CBL1. As funções aprende_parâmetros e aprende_parâmetros_em são rotinas que estimam os parâmetros usando o método de máxima verossimilhança.

1. $G := CBL1(D')$
2. $\theta := aprende\_parametros(G, D')$
3. loop
4. $\quad G_{novo} := CBL1(D' \cup D'')$
5. $\quad \theta := aprende\_parametros\_em(G_{novo}, D' \cup D'')$
6. $\quad$ if $G_{novo} = G$ // Se G_{novo} contém os mesmos arcos que G
7. $\quad\quad$ return $\langle G_{novo}, \theta \rangle$
8. $\quad$ else $G := G_{novo}$

Não existem equações fechadas para os estimadores que maximizam (1) e o logaritmo da somatória presente na equação dificulta sua maximização por métodos numéricos, por essa razão utiliza-se o algoritmo iterativo EM [Ghahramani e Jordan, 1994] que a cada passo aumenta o valor da verossimilhança e pára em um um máximo local.

A adaptação do algoritmo TAN para tratamento de dados não classificados foi feita em [Meila-Predoviciu, 1999]. Meila utiliza a propriedades específicas de árvores, relacionadas à verossimilhança, para desenvolver um algoritmo que aprende *simultaneamente* parâmetros e estrutura. A cada iteração do algoritmo EM uma nova estrutura é aprendida.

A nossa implementação do algoritmo CBL1 é baseado em um método iterativo que apresentou bons resultados na maior parte dos casos verificados. Em primeiro lugar, utilizamos o CBL1 para aprender uma rede a partir apenas dos dados classificados. Fixada a estrutura, o algoritmo EM é utilizado para aprender os parâmetros. Uma nova estrutura é aprendida porém dessa vez é utilizada a base classificada, e a não-classificada que agora conta com os valores dos "rótulos esperados". Então, o processo é repetido. O algoritmo pára quando duas estruturas consecutivas forem idênticas. A Tabela 1 apresenta uma visão esquemática do algoritmo proposto.

4. Experimentos

Nos experimentos realizados foram utilizadas quatro bases de dados provenientes do repositório da UCI [Blake e Merz, 1998] e duas bases de dados de reconhecimento de imagens. Os dados são provenientes de aplicações bastante diversas e foram escolhidas por conterem uma grande quantidade de exemplos, na ordem de milhares.

O algoritmo CBL1 requer a completa ordenação dos nós. Como essa informação é fornecida ao algoritmo na forma de conhecimento *a priori*, uma ordenação errada pode induzir o algoritmo a encontrar uma estrutura que não seja a correta, em geral com arcos adicionais. Por essa razão realizamos uma busca pela melhor ordenação. Apesar do espaço de busca ser $n!$, o fato de um grande número de ordenações serem equivalentes (especialmente se considerarmos estruturas esparsas) favorece a estratégia de busca sobre um espaço reduzido escolhido de forma aleatória. Para cada base de dados foram geradas 200 ordenações possíveis, para cada ordenação uma estrutura foi induzida e ordenada de acordo com a taxa de acerto sobre uma base de validação. No aprendizado semi-

Tabela 2: Resultados obtidos (taxa de acerto). Colunas: AT. contém o número de atributos; C. denota o número de exemplos classificados e NC. o número de não classificados. As abreviações C e NC também identificam o tipo de dado utilizado nos testes

	Descrição das bases				Resultados				
Bases de dados	AT.	Treino		Teste	base C.		base C. + NC.		
		C	NC.		NB	TAN	NB	TAN	CBL1
Satimage	37	600	3835	2000	81.6	83.5	77.5	81.0	83.5
Shuttle	10	100	43400	14500	82.4	81.2	76.1	90.5	91.8
Adult	13	6000	24163	15060	83.9	84.7	73.1	80.0	82.7
Chess	37	150	1980	1060	79.8	86.9	62.1	71.2	81.0
Cohn-Kanade	13	200	2980	1000	72.5	72.9	69.1	69.3	66.2
Chen-Huang	13	300	11982	3555	71.2	72.4	58.5	62.9	65.9

supervisionado dados completos são raros e não podem ser separados para compor uma base de validação e, por essa razão, utilizamos os dados classificados disponíveis na base de treino como base de validação. A melhor estrutura segundo esse critério foi utilizada para classificar a base de treino. Os testes de independência implementados, baseados em informação mútua, requerem a escolha de um *threshold*. Em todos os testes relatados, esse parâmetro esteve ajustado para 0.01.

Os resultados de classificação obtidos são reproduzidos na Tabela 2. A taxa de acerto é definida como a relação entre o número de predições corretas sobre o tamanho da base de teste. Os resultados obtidos são comparados com classificadores treinados apenas com a parcela de dados completos disponíveis.

Em cinco das seis bases de dados, o algoritmo modificado CBL1 teve um desempenho superior ao apresentado por NB e TAN (veja as três colunas à direita) sugerindo a eficácia da abordagem. Porém, na maior parte das bases estudadas a adição de exemplo não classificados, na maior parte dos casos, reduz a taxa de acerto do classificador. Essa aparente contradição evidencia a dificuldade de se tratar dados não classificados, o que é coerente com os resultados encontrados em trabalhos correlatos [Cozman e Cohen, 2002, Ghani, 2001].

5. Conclusões e Trabalhos Futuros

Neste trabalho foram apresentados os resultados obtidos com classificadores Bayesianos existentes para aprendizado semi-supervisionado, o NB e o TAN, além da utilização de um classificador baseado no algoritmo CBL1. Note que outros algoritmos de aprendizado, ao invés do CBL1, poderiam ser utilizados no processo iterativo apresentado na tabela 1 para tratar dados classificados e não classificados.

O algoritmo de aprendizado escolhido, o CBL1, requer uma ordenação completa das variáveis, o que traz benefícios práticos: simplifica o processo de aprendizado, pois menos testes de independência são necessários; é simples utilizá-lo para o aprendizado de classificadores baseados em estruturas generativas, pois basta definir a classe como sendo o primeiro nó na ordenação e permite a orientação de todos os arcos.

Comparando os algoritmos que tratam dados classificados e não classificados

(bases C + NC) verificamos que a versão modificada do CBL1 apresentou um desempenho superior ao NB e TAN. Esse resultado reforça a afirmação que a recuperação de relações de independência entre as variáveis é um caminho proveitoso para fazer uso dos dados não classificados disponíveis.

Apesar da aparente superioridade do CBL1, estruturas do tipo TAN devem ser sempre levadas em conta. Isso porque TAN tem complexidade $O(kn^2)$ (onde k é o número máximo de iterações do EM e n o número de variáveis) e as restrições estruturais impedem uma explosão do número de parâmetros. Mais importante, estruturas tipo TAN são especialmente úteis para o aprendizado semi-supervisionado em tarefas de reconhecimento de imagens.

6. Agradecimentos

Os autores agradecem ao Instituto Eldorado e à HP pela contribuição dada ao laboratório.

Referências

Baluja, S. (1999). Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *NIPS 11, 1999*, páginas 854–860.

Blake, C. e Merz, C. (1998). UCI repository of machine learning databases.

Cheng, J. e Greiner, R. (1999). Comparing Bayesian network classifiers. *UAI, 1999*.

Cheng, J., Bell, D. A., , e Liu, W. (1997). An algorithm for Bayesian network construction from data. *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*.

Chow, C. e Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Cohen, I., Sebe, N., Cozman, F. G., Cirelo, M. C., e Huang, T. S. (2003). Learning Bayesian network classifiers for facial expression recognition with both labeled and unlabeled data. *IEEE conference on Computer Vision and Pattern Recognition*.

Cowell, R. G. (2001). On searching for optimal classifiers among Bayesian networks. *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, páginas 175–180.

Cozman, F. G. e Cohen, I. (2002). Unlabeled data can degrade classification performance of generative classifiers. *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, páginas 327–331.

Domingos, P. e Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130.

Duda, R. e Hart, P. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Friedman, N., Geiger, D. e Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163.

Ghahramani, Z. e Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. *NIPS 7, 1994*, páginas 120–127.

Ghani, R. (2001). Combining labeled and unlabeled data for text classification with a large number of categories. *Proceedings of the First IEEE International Conference on Data Mining*, páginas 597–598.

Meila-Predoviciu, M. (1999). *Learning with Mixtures of Trees*. Tese de PhD, Massachusetts Institute of Technology.

Nigam, K., Mccallum, A. K., Thrun, S., e Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.