Biclusterização, dados faltantes e múltiplas imputações

Rosana Veroneze, Fernando J. Von Zuben

Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)

{veroneze, vonzuben}@dca.fee.unicamp.br

Abstract – Although the missing data problem has been studied for many years, it is still a relevant and challenging problem nowadays. There are several techniques capable of dealing with missing data. A parcel of them tries to estimate the missing values, which is called imputation. Nowadays, the most important technique is called Multiple Imputation (MI). Typically, MI adopts statistical concepts which assume that the model of data distribution is normal and that the mechanism associated with missing data is ignorable. This paper wants to show that the biclustering is an efficient and more flexible model to be used with MI. To achieve this objective, the quality of results was measured using the metrics described by Rubin in 1987. Moreover, the results applying biclustering were compared with the results of an MI classical program, the NORM. The experiments indicate that the biclustering was an efficient model, being better than NORM in all experiments.

Keywords – Missing data; data imputation; biclustering; multiple imputation.

1. Introdução

Com o avanço da tecnologia, a demanda por aquisição, armazenamento e processamento de uma enorme quantidade de dados é cada vez maior [2]. Com esse crescimento contínuo da disponibilidade de dados de diferentes fontes, surgem questões desafiadoras relacionadas à qualidade dos dados [18]. Um dos grandes problemas da maioria dos bancos de dados no mundo é a imprecisão e a incompletude, ou seja, a presença de dados ruidosos e faltantes, respectivamente [9]. Há muitas razões que levam a este cenário, entre elas: mau funcionamento de equipamentos, alto custo da coleta de dados e participantes de uma pesquisa que se recusam a responder a certas questões. Em muitas áreas, não é raro encontrar conjuntos de dados que tenham 50% ou mais de dados ruidosos ou faltantes [10].

Geralmente, três tipos de problemas estão associados aos dados faltantes: (*i*) perda de eficiência, (*ii*) restrições operacionais na manipulação e análise dos dados, e (*iii*) viés resultante das diferenças entre os valores atribuídos aos dados faltantes e os valores reais [10]. Um tratamento inadequado dos dados faltantes também pode afetar a generalização dos resultados obtidos [5, 12] Certamente, o melhor cenário seria evitar a ocorrência dos dados faltantes, e algumas estratégias podem ser adotadas para promover a disponibilidade de dados, como o aumento dos benefícios aos participantes de uma pesquisa [12], ou repetir o experimento [5]. No entanto, nem sempre é possível evitar os dados faltantes e, neste caso, não há outra forma de se obter resultados confiáveis, a não ser tratando os dados faltantes.

Existem vários métodos para o tratamento dos dados faltantes, dentre eles estão os métodos de imputação única, os métodos de máxima verossimilhança e os métodos de múltiplas imputações (MI). Os métodos de imputação única consistem, basicamente, em substituir os dados faltantes por um valor factível. Os métodos de máxima verossimilhança objetivam a estimação de parâmetros de modelos vinculados à distribuição dos dados. Enquanto os métodos de imputação única substituem cada valor faltante por um único valor, os métodos de MI substituem por x ($x \geq 2$) valores. Desse modo, são formadas x bases de dados completas que podem ser analisadas através de procedimentos convencionais. Os resultados dessas análises são agregados, gerando estimativas únicas para os parâmetros de interesse.

Ao contrário dos métodos de imputação única, os métodos de MI não tendem a subestimar a variabilidade da amostra [11]. Além disso, os métodos de MI são provavelmente menos sensíveis que os métodos de verossimilhança na escolha do modelo, porque o modelo é usado somente para a imputação dos valores e não para estimar os parâmetros [1]. Por isso, com o avanço da computação e com a proliferação de softwares que implementam algum método de MI, eles se tornaram rapidamente o método mais indicado para manipular dados faltantes [12].

Os softwares, que implementam MI utilizam procedimentos estatísticos para realizar as x imputações. Normalmente, são algoritmos que usam técnicas de Monte Carlo em cadeias de Markov [14], como o algoritmo chamado *Data Augmentation* (DA) [17]. Esses algoritmos requerem que sejam feitas, no mínimo, duas suposições. A primeira é sobre a distribuição dos

dados e a segunda é sobre o mecanismo dos dados faltantes. Como é difícil conhecer o mecanismo dos dados faltantes e a distribuição dos dados, principalmente em bases de dados incompletas, é interessante considerar técnicas de imputação que não dependam dessas informações, mas que, ao mesmo tempo, respeitem as relações existentes entre os dados observados da base de dados. Uma técnica capaz de extrair subconjuntos de linhas e colunas de uma matriz de dados, de modo que os elementos desses subconjuntos compartilhem alguma relação entre si, é a biclusterização [3].

Utilizar a biclusterização como modelo para as MIs tem suas vantagens. A primeira delas é não ser necessário considerar um modelo global para toda a base de dados. Uma vez que o processo de imputação ocorre apenas em um subconjunto menor e cuidadosamente selecionado de dados, a imputação sofre menos influência do ruído e da falta de outros dados, pois esses não participam de todo o processo de estimação. Além do mais, como os biclusters mostram explicitamente quais objetos e atributos foram selecionados, eles permitem interpretar e explicar diretamente os modelos e o mecanismo por trás da falta de dados.

Para trabalhar com as MIs juntamente com a biclusterização, foi utilizado um algoritmo de biclusterização, denominado SwarmBCluster, recentemente proposto por [6], que foi especializado para se tornar uma técnica de imputação de dados faltantes [7]. Ele foi escolhido por ser atualmente o único algoritmo de biclusterização que trabalha com a imputação de dados numéricos reais.

Esse artigo está organizado como segue. A seção 2 fala sobre as MIs e define as métricas utilizadas para medir a eficiência da biclusterização como modelo para as MIs. A seção 3 mostra os experimentos realizados com essa proposta. E, finalmente, na seção 4 alguns comentários finais e direções futuras da pesquisa serão delineados.

2. Múltiplas Imputações

Os métodos de múltiplas imputações (MI) foram propostos por [13]. Os métodos de MI geram x bases de dados completas, as quais são analisadas e os resultados são agregados.

Uma maneira simples para se gerar uma estimativa global para um parâmetro de interesse é através da média das estimativas produzidas para as x bases de dados [12]. Cada parâmetro de interesse estimado é chamado de $\hat{Q}$ e a estimativa global é chamada de $\overline{Q}$.

Já o cálculo do erro padrão global, que é necessário para os testes de significância e para os intervalos de confiança [12], não é tão trivial.

O procedimento para calcular o erro padrão global foi descrito por [13]. Primeiro é necessário calcular a variabilidade dos erros padrão que foram calculados para cada uma das x imputações, o que é chamado de *within-imputation variance*:

$$\overline{U} = \frac{1}{x}\sum_{j=1}^{x} U_j, \qquad (1)$$

onde U_j é o erro padrão ao quadrado da imputação j. Em seguida, é necessário calcular a variância de cada parâmetro estimado, o que é chamado de *between-imputation variance*:

$$B = \frac{1}{x-1}\sum_{j=1}^{x}\left(\hat{Q}_j - \overline{Q}\right)^2. \qquad (2)$$

Assim, é possível calcular a *variância total* de $\left(\hat{Q} - \overline{Q}\right)$:

$$T = \overline{U} + \left(1 + \frac{1}{x}\right)B. \qquad (3)$$

O erro padrão global é simplesmente a raiz quadrada de T.

Os graus de liberdade (df – do inglês *degrees of freedom*) são calculados através da Equação:

$$df = (x-1)\left(1 + r^{-1}\right)^2, \qquad (4)$$

onde r representa o aumento relativo na variância devido aos dados faltante e é dado por:

$$r = \frac{\left(1 + x^{-1}\right)B}{\overline{U}}. \qquad (5)$$

Quanto maior o valor de r, menor a estabilidade nos parâmetros estimados, refletindo em menor certeza estatística [12].

Para se calcular um intervalo com $100(1 - \alpha)\%$ de confiança em $\overline{Q}$, basta fazer:

$$\overline{Q} \pm t_{df}(\alpha/2)\sqrt{T}, \qquad (6)$$

onde $t_{df}(\alpha/2)$ é o extremo superior do intervalo obtido na Eq. 6, através da distribuição t de *Student* com df graus de liberdade. Por exemplo, se $df = \infty$ e $1 - \alpha = ,95 \rightarrow t_{df}(\alpha/2) = 1{,}96$, portanto para se calcular um intervalo de confiança de 95%, basta fazer: $\overline{Q} \pm 1{,}96\left(\sqrt{T}\right)$.

Um diferencial dos métodos de MI é que, além de estimarem os parâmetros de interesse e os erros padrão globais, eles também são capazes de estimar a incerteza estatística devido à presença dos dados faltantes. A taxa de informação faltante devido aos dados faltantes (γ) varia no intervalo [0, 1], sendo que 1 significa que existe 100% de informação faltante. Logo, quanto maior o valor de γ, menor a certeza estatística.

$$\gamma = \frac{r + 2/(df + 3)}{r + 1} \quad (7)$$

Rubin [13] também mostrou que a eficiência de uma estimativa baseada em x imputações em relação a uma baseada em um número infinito de imputações é:

$$\frac{1}{1 + \gamma / x} \quad (8)$$

Percebe-se que a eficiência relativa da inferência MI está relacionada com a taxa de informações faltantes (γ) em combinação com o número de imputações (x). Segundo [13], o valor de x pode ser restringido a um valor menor que 10. Além do mais, segundo [16], quando *df* é grande, a variância total é bem estimada e pouco se ganha em aumentar o valor de x.

3. Resultados

Os experimentos foram realizados em uma base de dados de expressão gênica bem conhecida, chamada popularmente de *Yeast* [4]. Essa base de dados contém 2884 genes (objetos) sob 17 condições (atributos) e pode ser encontrada em http://arep.med.harvard.edu/biclustering.

Como esses experimentos visam verificar o desempenho da biclusterização como modelo para os métodos de MI sob diferentes porcentagens de dados faltantes, foram geradas três bases de dados artificiais a partir da *Yeast*. Para isso, os dois objetos da base *Yeast* que possuem dados faltantes foram desconsiderados e, aleatoriamente, foram inseridos 15%, 50% e 80% de dados faltantes, segundo o mecanismo MCAR.

Os resultados obtidos pelo método de MI com a biclusterização foram comparados aos resultados obtidos pelo programa NORM [15]. O NORM é um programa gratuito para Windows, que cria MIs para bases de dados incompletas, considerando que os dados são distribuídos normalmente e que o mecanismo associado aos dados faltantes é ignorável. Ele utiliza dois algoritmos: o EM [8, 14] para gerar as estimativas iniciais dos parâmetros de interesse e o DA [17] para gerar as x imputações.

As Tabelas 1 e 2 trazem os resultados obtidos pela biclusterização e pelo NORM, respectivamente, para a base de dados de 15% de dados faltantes. O número de imputações foi igual a 5. Os parâmetros utilizados pelo SwarmBCluster foram: número de iterações = 1, número de formigas = 1, peso do feromônio = 1, peso da informação local = 3, taxa de evaporação do feromônio = 0,2, número mínimo de colunas de um bicluster = 7, número mínimo de linhas de um bicluster = 100, resíduo máximo de um bicluster = 200, número máximo de dados faltantes em um bicluster = 2.500. Os parâmetros utilizados pelo NORM foram: critério de convergência do EM: 1.00E-04, número de iterações do DA: 100.

Apesar do NORM ter obtido bons resultados para *df*, os resultados da biclusterização foram superiores, indicando que a variância total foi bem estimada. A biclusterização também obteve melhores resultados para *r*, o que indica maior estabilidade nos parâmetros estimados, consequentemente refletindo em maior certeza estatística (γ). A biclusterização também foi superior na eficiência das estimativas obtidas com 5 imputações comparadas a um número infinito de imputações (*Efic.*), o que indica que pouco se ganharia aumentando o número de imputações.

Para a base de dados com 50% de dados faltantes, os resultados foram equivalentes ao caso de 15% de dados faltantes. A biclusterização também obteve melhores resultados para *r*, refletindo em maior certeza estatística (γ).

Tabela 1. Biclusterização aplicada à base com 15% de dados faltantes.

Atr.	*df*	*r*	γ	*Efic.*
1	2142516,2429	0,0014	0,0014	0,9997
2	1296640,2415	0,0018	0,0018	0,9996
3	59810763,4941	0,0003	0,0003	0,9999
4	62467540,9174	0,0003	0,0003	0,9999
5	61211396,7743	0,0003	0,0003	0,9999
6	3309858,7342	0,0011	0,0011	0,9998
7	121920289,0703	0,0002	0,0002	1,0000
8	1879927,6659	0,0015	0,0015	0,9997
9	2707211,5139	0,0012	0,0012	0,9998
10	691012309,1395	0,0001	0,0001	1,0000
11	5120601,2548	0,0009	0,0009	0,9998
12	168338463,5425	0,0002	0,0002	1,0000
13	98203671,3355	0,0002	0,0002	1,0000
14	25364422,4771	0,0004	0,0004	0,9999
15	7863328,8928	0,0007	0,0007	0,9999
16	11646103,9076	0,0006	0,0006	0,9999
17	57117417,9223	0,0003	0,0003	0,9999

Tabela 2. NORM aplicado à base com 15% de dados faltantes.

Atr.	*df*	*r*	γ	*Efic.*
1	2163,2862	0,0449	0,0439	0,9913
2	111922,7329	0,0060	0,0060	0,9988
3	13077,6492	0,0178	0,0176	0,9965
4	562255,9246	0,0027	0,0027	0,9995
5	53131,4626	0,0088	0,0087	0,9983
6	64485,6486	0,0079	0,0079	0,9984
7	8918,6447	0,0216	0,0214	0,9957
8	17267,0252	0,0155	0,0153	0,9969
9	83493,0770	0,0070	0,0069	0,9986
10	367252,9673	0,0033	0,0033	0,9993
11	195273,1947	0,0045	0,0045	0,9991
12	157112,8434	0,0051	0,0051	0,9990
13	308087,3046	0,0036	0,0036	0,9993
14	28353,0342	0,0120	0,0119	0,9976
15	248935,7537	0,0040	0,0040	0,9992
16	432502,9766	0,0031	0,0030	0,9994
17	47201,9615	0,0093	0,0092	0,9982

As Tabelas 3 e 4 trazem os resultados obtidos pela biclusterização e pelo NORM, respectivamente, para a base de dados com 80% de dados faltantes. O número de imputações foi igual a 3, porque foi o máximo de imputações que o NORM conseguiu realizar. Os parâmetros utilizados pelo SwarmBCluster foram os mesmos dos experimentos anteriores, com exceção do resíduo máximo de um bicluster (500). Para o NORM, foi utilizado o mesmo critério de convergência do EM e o número de iterações do DA foi 450. Tanto a biclusterização como o NORM geraram estimativas com uma incerteza estatística consideravelmente maior que nos dois primeiros casos. Também, nesse experimento, a biclusterização não foi superior ao NORM em todos os atributos, como aconteceu nos dois primeiros casos. A grande vantagem da biclusterização, neste caso, é que, ao contrário do NORM, ela poderia trabalhar com um número de imputações maior, o que implicaria em maior certeza estatística.

Tabela 3. Biclusterização aplicada à base com 80% de dados faltantes.

Atr.	*df*	*r*	*γ*	*Efic.*
1	10,5707	0,7698	0,5182	0,8527
2	399,8189	0,0761	0,0753	0,9755
3	23,7477	0,4089	0,3433	0,8973
4	539,6932	0,0648	0,0643	0,9790
5	1400,7913	0,0393	0,0392	0,9871
6	12,8727	0,6506	0,4705	0,8644
7	192,3948	0,1135	0,1111	0,9643
8	1354,5162	0,0400	0,0398	0,9869
9	129,2494	0,1421	0,1376	0,9561
10	4,8218	1,8093	0,7351	0,8032
11	32,0431	0,3330	0,2926	0,9111
12	149,7994	0,1306	0,1271	0,9593
13	442,8922	0,0720	0,0714	0,9768
14	140,0720	0,1357	0,1318	0,9579
15	191514,2577	0,0032	0,0032	0,9989
16	88,4831	0,1769	0,1689	0,9467
17	412,4276	0,0748	0,0741	0,9759

Tab. 4. NORM aplicado à base com 80% de dados faltantes.

Atr.	*df*	*r*	*γ*	*Efic.*
1	12,0488	0,6875	0,4862	0,8605
2	212,4582	0,1074	0,1054	0,9661
3	107,3314	0,1581	0,1522	0,9517
4	30,6629	0,3430	0,2996	0,9092
5	7,7439	1,0333	0,5997	0,8334
6	59,5382	0,2244	0,2094	0,9348
7	9,1642	0,8767	0,5548	0,8439
8	20,0863	0,4610	0,3748	0,8889
9	32,7437	0,3283	0,2893	0,9121
10	5,3602	1,5696	0,7039	0,8099
11	18,3495	0,4929	0,3929	0,8842
12	2543,7096	0,0288	0,0288	0,9905
13	13,4946	0,6260	0,4595	0,8672
14	223,8428	0,1044	0,1025	0,9670
15	27252,1555	0,0086	0,0086	0,9971
16	393,0909	0,0768	0,0760	0,9753
17	20,5073	0,4541	0,3708	0,8900

4. Conclusões

Esse artigo investigou o comportamento da biclusterização como modelo para as MIs. Ela se mostrou um modelo eficiente, superando os resultados obtidos por um programa clássico de MIs, o NORM. Como trabalhos futuros, deve-se levar em conta diferentes bases de dados e diferentes mecanismos de dados faltantes.

Referências

[1] P. D. ALLISON. *Missing data***.** Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage, 2001.

[2] M. L. BROWN, J. F. KROS. Data mining and the impact of missing data. *Industrial Management & Data Systems***,** p. 611-621, 2003.

[3] Y. CHENG, G. M. CHURCH. Biclustering of expression data. *Proceedings of the 8th Int. Conf. on Intelligent Systems for Molecular Biology*, p. 93–103, 2000.

[4] R. CHO, M. CAMPBELL, E. WINZELER, L. STEINMETZ, A. CONWAY, L. WODICKA, A. GABRIELIAN, D. LANDSMAN, D. LOCKHART, R. DAVIS. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, v. 2, n. 1, p. 65–73, 1998.

[5] COLANTONIO, R. D. PIETRO, A. OCELLO, N. V. VERDE. ABBA: Adaptive Bicluster-Based Approach to Impute Missing Values in Binary Matrices. *Proceedings of the 2010 ACM Symposium on Applied Computing*, p. 1026-1033, 2010.

[6] F. O. DE FRANÇA, F. J. VON ZUBEN. Finding a high coverage set of δ-biclusters with swarm intelligence. *Proceedings of the 12th IEEE Congress on Evolutionary Computation*, p. 1 – 8, 2010.

[7] F. O. DE FRANÇA. *Biclusterização na Análise de Dados Incertos*. Tese de Doutorado, UNICAMP, Campinas, 2010.

[8] A. P. DEMPSTER, N. M. Laird, D. B. RUBIN. Maximum-likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1-38, 1977.

[9] FARHANGFAR, L. KURGAN, W. PEDRYCZ. Experimental analysis of methods for imputation of missing values in databases. *Proceedings of SPIE*, v. 5421, p. 172–182, 2004.

[10] FARHANGFAR, L. KURGAN, W. PEDRYCZ. A Novel Framework for Imputation of Missing Values in Databases. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, v. 37, n. 5, p. 692-709, 2007.

[11] R. J. A. LITTLE, D. B. RUBIN. *Statistical analysis with missing data*. 2. ed. Hoboken: John Wiley & Sons, 2002.

[12] P. E. MCKNIGHT, K. M. MCKNIGHT, S. SIDANI, A. J. FIGUEREDO. *Missing data: a gentle introduction*. New York: The Guilford Press, 2007.

[13] D. B. RUBIN. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[14] J. L. SCHAFER. *Analysis of Incomplete Multivariate Data*. Londom: Chapman & Hall, 1997.

[15] J. L. SCHAFER. (1999) *NORM*: *Multiple imputation of incomplete multivariate data under a normal model*, version 2. Programa para Windows 95/98/NT, disponível em http://www.stat.psu.edu/~jls/misoftwa.html.

[16] J. L. SCHAFER, J. W. GRAHAM. Missing Data: Our View of the State of the Art. *Psychological Methods*, v. 7, n. 2, p. 147-177, 2002.

[17] M. A. TANNER, W. H. WONG. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550, 1987.

[18] WU, C. WUN, H. CHOU. Using association rules for completing missing data. *Fourth International Conference On Hybrid Intelligent Systems*, Taiwan, p. 236-241, 2004.