Recuperação inteligente da informação
Relatório atividade III

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

Conteúdo

1 Introdução

A terceira atividade nada mais é do que uma continuação das duas atividades anteriores. Utilizando a matriz composta pelos valores de todas as métricas, o intuito desta atividade é comparar cada uma as mesmas de acordo com o coeficiente de correlação entre elas. O coeficiente de correlação, como o próprio nome sugere, tem como objetivo medir o quão relacionadas estão duas variáveis. Portanto, neste relatório são apresentados todos os valores de correlação para cada comparação par a par e uma conclusão em relação as métricas de acordo com esses valores.

2 Coeficiente de correlação

O coeficiente de correlação, normalmente representado por ρ, é calculado da seguinte forma:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X) \times var(Y)}} \tag{1}$$

sendo X e Y duas variáveis com mesmo número de dados, $cov(X,Y)$ e $var(X,Y)$ a covariância e variância entre as variáveis, respectivamente. O valor do coeficiente pode assumir valores no intervalo $-1 \leq \rho \leq 1$. Se $\rho = 0$, o mesmo indica que

as variáveis não possuem correlação. Por outro lado, se $\rho = 1$ ou $\rho = -1$, o coeficiente indica que as variáveis são perfeitamente correlacionadas, positivamente e negativamente, respectivamente.

Na atividade, são disponibilizadas 37 métricas com 17 valores para cada uma delas. Neste caso, seriam necessários combinar as 37 variáveis par a par sem repetição para o cálculo de cada uma das correlações. Para facilitar esse cálculo, todos os valores são organizados em uma matriz e calcula-se a matriz de correlação. Essa matriz é composta por todos os coeficientes de correlação como descrito a seguir:

$$\begin{bmatrix} \rho(X_1, X_1) & \cdots & \rho(X_1, X_m) \\ \vdots & \ddots & \vdots \\ \rho(X_1, X_m) & \cdots & \rho(X_m, X_m) \end{bmatrix} \tag{2}$$

Sendo m o número de variáveis a ser comparadas. Vale a pena ressaltar que a diagonal da matriz de correlação é igual a 1 (pois a correlação da própria variável com ela mesmo é sempre 1) e a matriz triangular inferior é sempre igual a superior, pois a comparações se repetem.

3 Comparação entre as métricas

A matriz de correlação com todos os coeficientes de correlação entre as variáveis, por motivos de visualização, foi dividida e apresentada nas tabelas 1, 2 e 3. Além disso, para facilitar a distinção das comparações, a matriz triangular superior foi retirada, uma vez que todas as comparações já são apresentadas na matriz triangular inferior. Nas tabelas, as abreviações M1-37 substituem os nomes das métricas em ordem alfabética dos arquivos disponibilizados, como descrito na tabela 4.

De acordo com os valores descritos na tabela, podemos perceber que de maneira geral, a maioria das variáveis possuem um grau de correlação elevado. Além disso, a maior parte das correlações são positivas, ou seja, quando uma métrica cresce a outra também tende a crescer. Em poucos casos as variáveis possuem um valor muito baixo de correlação, como por exemplo as comparações M30-M10, M11-M8 e M26-M8.

Algumas métricas, como a M1, M2, M3 e M4, possuem, na maior parte, altas correlações positivas com todas as demais. Por outro lado, a métrica M34 possui muitas correlações negativas com as demais, ou seja, quando ela decresce, as demais tendem a crescer. Vale destacar algumas métricas perfeitamente correlacionadas, como os pares M35-M29 e M18-17. Neste último caso, os pares de variáveis tomam os mesmos valores.

Sendo assim, considerando a análise realizada, pode-se concluir que de maneira geral as métricas são redundantes entre si. Não existe a necessidade de se utilizar todas elas em um caso de comparação de desempenho de algoritmos, por exemplo. Neste caso, seria interessante descartar boa parte das mesmas.

Tabela 1. Matriz de correlação das métricas - parte 1

-	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**	**M13**
M1	1.000												
M2	0.644	1.000											
M3	0.922	0.629	1.000										
M4	0.645	0.425	0.531	1.000									
M5	0.177	0.706	0.147	0.148	1.000								
M6	0.762	0.861	0.691	0.540	0.484	1.000							
M7	0.310	0.286	0.427	0.335	-0.021	0.350	1.000						
M8	0.426	0.228	0.509	-0.229	0.021	0.208	0.125	1.000					
M9	0.657	0.145	0.722	0.330	-0.211	0.285	0.238	0.610	1.000				
M10	0.421	0.099	0.447	-0.021	-0.091	0.078	-0.034	0.708	0.531	1.000			
M11	0.729	0.368	0.748	0.753	-0.036	0.528	0.428	-0.035	0.466	0.029	1.000		
M12	0.429	0.761	0.404	0.255	0.814	0.535	-0.023	0.096	0.127	0.087	0.161	1.000	
M13	0.762	0.861	0.691	0.540	0.484	1.000	0.350	0.208	0.285	0.078	0.528	0.535	1.000
M14	0.645	0.473	0.698	0.189	0.141	0.474	0.246	0.715	0.736	0.714	0.222	0.398	0.474
M15	0.657	0.145	0.722	0.330	-0.211	0.285	0.238	0.610	1.000	0.531	0.466	0.127	0.285
M16	0.948	0.600	0.863	0.711	0.120	0.738	0.332	0.328	0.637	0.326	0.678	0.377	0.738
M17	0.932	0.596	0.862	0.735	0.133	0.736	0.346	0.323	0.630	0.317	0.683	0.355	0.736
M18	0.932	0.596	0.862	0.735	0.133	0.736	0.346	0.323	0.630	0.317	0.683	0.355	0.736
M19	0.861	0.484	0.898	0.676	0.047	0.603	0.540	0.440	0.673	0.494	0.766	0.192	0.603
M20	0.855	0.772	0.809	0.350	0.307	0.846	0.263	0.463	0.446	0.327	0.560	0.510	0.846
M21	0.402	-0.024	0.376	0.577	-0.101	0.121	0.389	0.263	0.454	0.293	0.363	-0.145	0.121
M22	0.942	0.699	0.851	0.716	0.277	0.746	0.283	0.274	0.527	0.272	0.719	0.536	0.746
M23	0.712	0.857	0.636	0.564	0.492	0.989	0.353	0.159	0.245	0.033	0.488	0.522	0.989
M24	0.494	0.370	0.653	0.635	0.148	0.356	0.385	0.021	0.320	0.168	0.687	0.160	0.356
M25	0.770	0.456	0.653	0.662	0.159	0.578	0.322	0.312	0.622	0.339	0.648	0.347	0.578
M26	0.494	0.370	0.653	0.635	0.148	0.356	0.385	0.021	0.320	0.168	0.687	0.160	0.356
M27	0.784	0.520	0.877	0.455	0.046	0.637	0.695	0.448	0.635	0.434	0.707	0.226	0.637
M28	0.773	0.409	0.819	0.356	-0.074	0.540	0.227	0.571	0.867	0.536	0.611	0.289	0.540
M29	0.481	0.706	0.419	0.549	0.502	0.827	0.322	0.096	0.063	-0.029	0.302	0.310	0.827
M30	0.762	0.861	0.691	0.540	0.484	1.000	0.350	0.208	0.285	0.078	0.528	0.535	1.000
M31	0.784	0.520	0.877	0.455	0.046	0.637	0.695	0.448	0.635	0.434	0.707	0.226	0.637
M32	0.912	0.643	0.997	0.541	0.171	0.699	0.439	0.503	0.715	0.445	0.735	0.411	0.699
M33	0.792	0.850	0.795	0.620	0.510	0.773	0.286	0.214	0.477	0.194	0.621	0.789	0.773
M34	-0.200	0.269	-0.333	0.180	0.454	0.127	-0.132	-0.715	-0.726	-0.535	0.003	0.287	0.127
M35	0.481	0.706	0.419	0.549	0.502	0.827	0.322	0.096	0.063	-0.029	0.302	0.310	0.827
M36	0.932	0.596	0.862	0.735	0.133	0.736	0.346	0.323	0.630	0.317	0.683	0.355	0.736
M37	0.471	0.371	0.395	0.911	0.260	0.418	0.312	-0.196	0.147	-0.069	0.638	0.176	0.418

Tabela 2. Matriz de correlação das métricas - parte 2

	M14	**M15**	**M16**	**M17**	**M18**	**M19**	**M20**	**M21**	**M22**	**M23**	**M24**	**M25**	**M26**
M1													
M2													
M3													
M4													
M5													
M6													
M7													
M8													
M9													
M10													
M11													
M12													
M13													
M14	1.000												
M15	0.736	1.000											
M16	0.652	0.637	1.000										
M17	0.652	0.630	0.994	1.000									
M18	0.652	0.630	0.994	1.000	1.000								
M19	0.593	0.673	0.804	0.823	0.823	1.000							
M20	0.627	0.446	0.770	0.737	0.737	0.615	1.000						
M21	0.262	0.454	0.487	0.531	0.531	0.648	0.007	1.000					
M22	0.517	0.527	0.895	0.885	0.885	0.803	0.778	0.349	1.000				
M23	0.444	0.245	0.710	0.716	0.716	0.580	0.772	0.152	0.718	1.000			
M24	0.253	0.320	0.498	0.558	0.558	0.749	0.224	0.506	0.531	0.373	1.000		
M25	0.454	0.622	0.698	0.669	0.669	0.705	0.639	0.496	0.690	0.531	0.261	1.000	
M26	0.253	0.320	0.498	0.558	0.558	0.749	0.224	0.506	0.531	0.373	1.000	0.261	1.000
M27	0.606	0.635	0.704	0.693	0.693	0.874	0.710	0.394	0.654	0.583	0.562	0.690	0.562
M28	0.706	0.867	0.658	0.636	0.636	0.733	0.702	0.222	0.668	0.490	0.341	0.706	0.341
M29	0.297	0.063	0.555	0.597	0.597	0.480	0.490	0.370	0.475	0.873	0.421	0.386	0.421
M30	0.474	0.285	0.738	0.736	0.736	0.603	0.846	0.121	0.746	0.989	0.356	0.578	0.356
M31	0.606	0.635	0.704	0.693	0.693	0.874	0.710	0.394	0.654	0.583	0.562	0.690	0.562
M32	0.703	0.715	0.861	0.864	0.864	0.905	0.788	0.401	0.841	0.652	0.684	0.640	0.684
M33	0.569	0.477	0.767	0.751	0.751	0.637	0.779	0.171	0.828	0.747	0.477	0.701	0.477
M34	-0.565	-0.726	-0.229	-0.231	-0.231	-0.284	-0.136	-0.374	0.004	0.180	0.032	-0.204	0.032
M35	0.297	0.063	0.555	0.597	0.597	0.480	0.490	0.370	0.475	0.873	0.421	0.386	0.421
M36	0.652	0.630	0.994	1.000	1.000	0.823	0.737	0.531	0.885	0.716	0.558	0.669	0.558
M37	0.033	0.147	0.522	0.568	0.568	0.605	0.200	0.666	0.565	0.450	0.665	0.568	0.665

Tabela 3. Matriz de correlação das métricas - parte 3

	M27	**M28**	**M29**	**M30**	**M31**	**M32**	**M33**	**M34**	**M35**	**M36**	**M37**
M1											
M2											
M3											
M4											
M5											
M6											
M7											
M8											
M9											
M10											
M11											
M12											
M13											
M14											
M15											
M16											
M17											
M18											
M19											
M20											
M21											
M22											
M23											
M24											
M25											
M26											
M27	1.000										
M28	0.754	1.000									
M29	0.397	0.187	1.000								
M30	0.637	0.540	0.827	1.000							
M31	1.000	0.754	0.397	0.637	1.000						
M32	0.878	0.805	0.455	0.699	0.878	1.000					
M33	0.651	0.650	0.520	0.773	0.651	0.797	1.000				
M34	-0.313	-0.469	0.177	0.127	-0.313	-0.321	0.042	1.000			
M35	0.397	0.187	1.000	0.827	0.397	0.455	0.520	0.177	1.000		
M36	0.693	0.636	0.597	0.736	0.693	0.864	0.751	-0.231	0.597	1.000	
M37	0.329	0.151	0.587	0.418	0.329	0.413	0.485	0.212	0.587	0.568	1.000

Tabela 4. Relação entre a a sigla e as métricas

M1	additivesymmetric
M2	average
M3	bhattacharrya
M4	canberra
M5	chebyshev
M6	cityblock
M7	clark
M8	cosseno
M9	czekanowski
M10	dice
M11	divergence
M12	euclidiana
M13	gower
M14	harmonicmean
M15	intersection
M16	jeffreys
M17	jensendifference
M18	jensenshannon
M19	kdivergence
M20	kulczynski
M21	kullbackIiebler
M22	kumarjohnson
M23	lorentzian
M24	matusita
M25	neyman
M26	nhllinger
M27	probabilisticsymmetric
M28	ruzicka
M29	soergel
M30	sorensen
M31	squaredchord
M32	squaredeuclidean
M33	squared
M34	taneja
M35	tanimoto
M36	topsoe
M37	wavehedges