Simulation study of a tree similarity measure based on small subtree counts

Mindaugas BLOZNELIS, Irmantas RADAVIČIUS
Vilnius University, Faculty of Mathematics and Informatics
Naugarduko 24, LT-03225 Vilnius, Lithuania
e-mail: mindaugas.bloznelis@mif.vu.lt; irmantas.radavicius@mif.vu.lt

Abstract. Augsten, Böhlen and Gamper [1] suggested a measure of similarity between ordered and labeled trees based on subtree counts: two trees are declared close if they contain similar number of copies of ordered and labeled subtrees of a given form, called pq-gram. We report the results of a simulation study of statistical properties of distances based on pq-grams.

Keywords: tree comparison, Galton–Watson tree.

1. Introduction

Augsten, Böhlen and Gamper [1] suggested a measure of similarity between (ordered and labeled) trees based on subtree counts. Roughly speaking, two trees are declared close if they contain similar numbers of copies of (ordered and labeled) subtrees of a given form, called pq-gram. Augsten, Böhlen and Gamper [1] used distances based on pq-grams to define approximate matching of hierarchical data.

Here we are interested in statistical properties of distances based on pq-grams. For this purpose we consider several parametric families of Galton–Watson random trees. Computer simulation results show that pq-gram distance effectively discriminates between two Galton–Watson trees generated for different values of parameter.

2. pg-gram distances of Galton–Watson trees

2.1. A rooted unlabeled tree on $p+q$ vertices is called pq-gram if for every $0 \leqslant j \leqslant p-1$ there is only one vertex in the distance j from the root and the vertex in the distance $p-1$ from the root has q leaves of degree 1 (which are in the distance p from the root), see Augsten, Böhlen and Gamper [1]. That is, in order to obtain the pq-gram we stick the (center of the) star $K_{1,q}$ to one endpoint of the path on p vertices. Another endpoint of the path is called the root of the pq-gram. The pq-gram is denoted $T^{p,q}$.

Using letters from an alphabet, say Σ, of size k we obtain $n = k^{p+q}$ labeled ordered pq-grams $T_1^{p,q}, \dots, T_n^{p,q}$. Given an ordered labeled tree T, with labels from Σ, we prescribe the vector $N(T) = (N_1, \dots, N_n)$ that counts copies of $T_i^{p,q}$, $1 \leqslant i \leqslant n$, contained in T. More precisely, N_i denotes the number of exact matchings of $T_i^{p,q}$ in T (order is important). One would expect that two ordered and labeled trees T_1, T_2 are similar if the corresponding subtree count vectors $N(T_1)$ and $N(T_2)$ were close.

Augsten, Böhlen and Gamper [1] define the pq-gram distance as follows. Let T be an ordered labeled tree with labels from the alphabet Σ. Introduce an extra letter $'*'$ and extend the alphabet $\Sigma^* = \Sigma \cup \{*\}$. The pq-extended tree T^* is constructed from T by adding $p-1$ ancestors to the root node, inserting $q-1$ children before the first and after the last child of each non-leaf node, and adding q children to each leaf of T. All newly inserted nodes become labels $'*'$. Let the subtree count vector $N_*(T^*) = (N_1^*(T^*), \dots, N_m^*(T^*))$ be defined as above but for the alphabet Σ^* and the extended tree T^*. In particular, we have $m = (k+1)^{p+q}$. The pq-gram distance between two ordered labeled trees T_1 and T_2

$$\Delta^{p,q}(T_1, T_2) = 1 - 2\frac{\sum_{i=1}^m \min\{N_i^*(T_1^*),\ N_i^*(T_2^*)\}}{\sum_{i=1}^m (N_i^*(T_1^*) + N_i^*(T_2^*))}.$$

2.2. One would expect that graph similarity measure based on small subgraph count should discriminate between graphs generated using different probabilistic models. Galton–Watson tree is a convenient probabilistic model to test statistical properties of the similarity measure based on pq-gram counts. Below we refer results of a simulation study. Given tree T, we denote by T^k the subtree induced by vertices that are in a distance of at most k from the root.

In Examples 1–3 we put $\Sigma = \{a\}$. Therefore, we have $\Sigma^* = \{a, *\}$. Every node of T^k is labeled with the letter a, while some nodes of its extended version receive also labels $'*'$.

Example 1. Given p we generate Galton–Watson tree $T(p)$ with binomial $Bi(10, p)$ offspring distribution. Table 1 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^7(p), T^7(p'))$, for $p, p' \in \{0.1, 0.2, \dots, 0.9\}$. These estimated values are based on computer simulation of 100 independent copies of $T(p)$ for each p.

Example 2. Given λ we generate Galton–Watson tree $T(\lambda)$ with Poisson offspring distribution with mean λ. Table 2 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^7(\lambda), T^7(\lambda'))$, for $\lambda, \lambda' \in \{0.125, 0.25, \dots, 8\}$. These estimated values are based on computer simulation of 100 independent copies of $T(\lambda)$ for each λ.

Example 3. Given p we generate Galton–Watson tree $T(p)$ with Geometric offspring distribution with parameter $p \in (0,1)$. Table 3 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^{100}(p), T^{100}(p'))$, for $p, p' \in \{2^{-1}, 2^{-2}, \dots, 2^{-7}\}$. These estimated values are based on computer simulation of 100 independent copies of $T(p)$ for each p.

In Examples 4–6 we put $\Sigma = \{a, b\}$. Therefore, we have $\Sigma^* = \{a, b, *\}$. We generate Galton–Watson tree T with two types of offspring a and b. Given a vertex of type a (respectively b) let X_{aa} and X_{ab} (respectively X_{ba} and X_{bb}) denote its offspring numbers of types a and b. Random variables $X_{aa}, X_{ab}, X_{ba}, X_{bb}$ are independent and have Poisson distributions with mean values $\lambda_{aa} = 5p(a|a)$, $\lambda_{ab} = 5p(b|a)$, $\lambda_{ba} = 5p(a|b)$, $\lambda_{aa} = 5p(b|b)$. We denote $T = T(\overline{p})$, where $\overline{p} = (p(a|a), p(b|a), p(a|b), p(b|b))$. The root of T chooses its label (a or b) at random and with equal probabilities.

Example 4. Here we choose $p(a|a) = p(a|b) = 1 - p(b|a) = 1 - p(b|b)$ where $p(a|a) \in \{0.1, 0.2, \dots, 0.9\}$. In this way we obtain 9 different vectors $\overline{p}_1, \dots, \overline{p}_9$. Table 4 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^{100}(\overline{p}_i), T^{100}(\overline{p}_j))$, for $1 \leqslant i, j \leqslant 9$. These estimated values are based on computer simulation of 100 independent copies of $T(\overline{p})$ for each $\overline{p}$.

Table 1. Results of simulation which was described in Example 1

p	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	0.1306	0.2202	0.3497	0.4259	0.4819	0.5268	0.5635	0.5928	0.6164
0.2	0.2202	0.0415	0.1429	0.2442	0.3190	0.3740	0.4148	0.4462	0.4710
0.3	0.3497	0.1429	0.0091	0.1033	0.1781	0.2331	0.2739	0.3053	0.3301
0.4	0.4259	0.2442	0.1033	0.0027	0.0749	0.1299	0.1707	0.2021	0.2269
0.5	0.4819	0.3190	0.1781	0.0749	0.0009	0.0550	0.0958	0.1272	0.1520
0.6	0.5268	0.3740	0.2331	0.1299	0.0550	0.0004	0.0408	0.0722	0.0970
0.7	0.5635	0.4148	0.2739	0.1707	0.0958	0.0408	0.0002	0.0314	0.0561
0.8	0.5928	0.4462	0.3053	0.2021	0.1272	0.0722	0.0314	0.0001	0.0248
0.9	0.6164	0.4710	0.3301	0.2269	0.1520	0.0970	0.0561	0.0248	0.0000

Table 2. Results of simulation which was described in Example 2

λ	**0.125**	**0.250**	**0.500**	**0.750**	**1.000**	**1.500**	**2.000**	**4.000**	**8.000**
0.125	0.1277	0.1939	0.3174	0.3948	0.4525	0.4993	0.5360	0.5651	0.5886
0.250	0.1939	0.0461	0.1406	0.2358	0.3061	0.3587	0.3986	0.4296	0.4543
0.500	0.3174	0.1406	0.0112	0.0981	0.1690	0.2216	0.2616	0.2925	0.3172
0.750	0.3948	0.2358	0.0981	0.0033	0.0709	0.1236	0.1635	0.1945	0.2192
1.000	0.4525	0.3061	0.1690	0.0709	0.0013	0.0527	0.0926	0.1236	0.1482
1.500	0.4993	0.3587	0.2216	0.1236	0.0527	0.0006	0.0399	0.0709	0.0956
2.000	0.5360	0.3986	0.2616	0.1635	0.0926	0.0399	0.0003	0.0310	0.0557
4.000	0.5651	0.4296	0.2925	0.1945	0.1236	0.0709	0.0310	0.0002	0.0247
8.000	0.5886	0.4543	0.3172	0.2192	0.1482	0.0956	0.0557	0.0247	0.0001

Table 3. Results of simulation which was described in Example 3

p	$\mathbf{(1/2)^1}$	$\mathbf{(1/2)^2}$	$\mathbf{(1/2)^3}$	$\mathbf{(1/2)^4}$	$\mathbf{(1/2)^5}$	$\mathbf{(1/2)^6}$	$\mathbf{(1/2)^7}$
$\mathbf{(1/2)^1}$	0.6323	0.6841	0.7771	0.8552	0.8934	0.9332	0.9402
$\mathbf{(1/2)^2}$	0.6841	0.5476	0.5732	0.6627	0.7325	0.7733	0.7945
$\mathbf{(1/2)^3}$	0.7771	0.5732	0.3561	0.3952	0.4770	0.5066	0.5369
$\mathbf{(1/2)^4}$	0.8552	0.6627	0.3952	0.1895	0.2541	0.2726	0.3091
$\mathbf{(1/2)^5}$	0.8934	0.7325	0.4770	0.2541	0.1526	0.1527	0.1896
$\mathbf{(1/2)^6}$	0.9332	0.7733	0.5066	0.2726	0.1527	0.0571	0.0882
$\mathbf{(1/2)^7}$	0.9402	0.7945	0.5369	0.3091	0.1896	0.0882	0.0699

Table 4. Results of simulation which was described in Example 4

$p(a\|a)$	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	0.0417	0.2584	0.4184	0.5518	0.6715	0.7672	0.8452	0.9107	0.9612
0.2	0.2584	0.0251	0.1945	0.3580	0.4965	0.6172	0.7407	0.8324	0.9100
0.3	0.4184	0.1945	0.0265	0.1794	0.3311	0.4840	0.6211	0.7385	0.8427
0.4	0.5518	0.3580	0.1794	0.0282	0.1764	0.3338	0.4904	0.6191	0.7663
0.5	0.6715	0.4965	0.3311	0.1764	0.0275	0.1748	0.3378	0.4972	0.6689
0.6	0.7672	0.6172	0.4840	0.3338	0.1748	0.0276	0.1890	0.3604	0.5494
0.7	0.8452	0.7407	0.6211	0.4904	0.3378	0.1890	0.0460	0.2036	0.4198
0.8	0.9107	0.8324	0.7385	0.6191	0.4972	0.3604	0.2036	0.0254	0.2511
0.9	0.9612	0.9100	0.8427	0.7663	0.6689	0.5494	0.4198	0.2511	0.0219

Table 5. Results of simulation which was described in Example 5

$p(a\|a)$	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	0.0478	0.2536	0.4017	0.5108	0.5936	0.6645	0.7158	0.7542	0.7941
0.2	0.2536	0.0287	0.1773	0.3021	0.3934	0.4685	0.5353	0.6016	0.6477
0.3	0.4017	0.1773	0.0472	0.1559	0.2478	0.3314	0.4174	0.4840	0.5372
0.4	0.5108	0.3021	0.1559	0.0476	0.1254	0.2172	0.3046	0.3745	0.4342
0.5	0.5936	0.3934	0.2478	0.1254	0.0280	0.1054	0.1956	0.2719	0.3404
0.6	0.6645	0.4685	0.3314	0.2172	0.1054	0.0280	0.1060	0.1827	0.2629
0.7	0.7158	0.5353	0.4174	0.3046	0.1956	0.1060	0.0460	0.1137	0.1928
0.8	0.7542	0.6016	0.4840	0.3745	0.2719	0.1827	0.1137	0.0440	0.1103
0.9	0.7941	0.6477	0.5372	0.4342	0.3404	0.2629	0.1928	0.1103	0.0234

Table 6. Results of simulation which was described in Example 6

$p(a\|a)$	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	0.0502	0.2314	0.3856	0.5153	0.6125	0.7138	0.7942	0.8651	0.9080
0.2	0.2314	0.0539	0.1840	0.3393	0.4603	0.5838	0.6728	0.7479	0.8675
0.3	0.3856	0.1840	0.0301	0.1785	0.3092	0.4368	0.5280	0.6728	0.7980
0.4	0.5153	0.3393	0.1785	0.0672	0.1789	0.3089	0.4376	0.5844	0.7193
0.5	0.6125	0.4603	0.3092	0.1789	0.0470	0.1781	0.3091	0.4594	0.6192
0.6	0.7138	0.5838	0.4368	0.3089	0.1781	0.0658	0.1791	0.3399	0.5245
0.7	0.7942	0.6728	0.5280	0.4376	0.3091	0.1791	0.0256	0.1830	0.3976
0.8	0.8651	0.7479	0.6728	0.5844	0.4594	0.3399	0.1830	0.0425	0.2457
0.9	0.9080	0.8675	0.7980	0.7193	0.6192	0.5245	0.3976	0.2457	0.0588

Example 5. Here we choose $p(a|a)=1-p(b|a)$ where $p(a|a)\in\{0.1,0.2,\dots,0.9\}$, and $p(a|b)=p(b|b)=0.5$. In this way we obtain 9 different vectors $\overline{p}_1,\dots,\overline{p}_9$. Table 5 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^{100}(\overline{p}_i),T^{100}(\overline{p}_j))$, for $1\leqslant i,j\leqslant 9$. These estimated values are based on computer simulation of 100 independent copies of $T(\overline{p})$ for each $\overline{p}$.

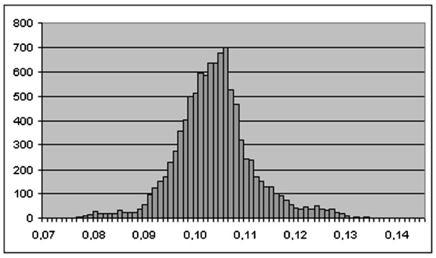

Fig. 1. Histogram from Example 7.

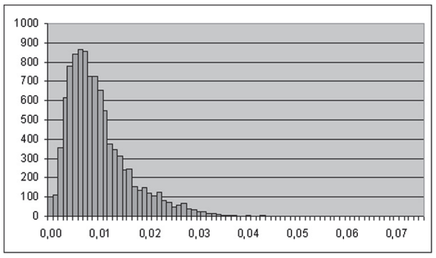

Fig. 2. Histogram from Example 7.

Example 6. Here we choose $p(a|a) = 1 - p(b|a) = p(b|b) = 1 - p(a|b)$ where $p(a|a) \in \{0.1, 0.2, \ldots, 0.9\}$. In this way we obtain 9 different vectors $\overline{p}_1, \ldots, \overline{p}_9$. Table 6 shows estimated values of mathematical expectations of 23-gram distances $\Delta^{2,3}(T^{100}(\overline{p}_i), T^{100}(\overline{p}_j))$, for $1 \leqslant i, j \leqslant 9$. These estimated values are based on computer simulation of 100 independent copies of $T(\overline{p})$ for each $\overline{p}$.

Example 7. We study the value distribution of $\Delta^{2,3}(T^7(p), T^7(p'))$ defined in Example 1. Fig. 1 shows the histogram of the value distribution in the case where $p = 0.3$ and $p' = 0.4$. Fig. 2 shows the histogram of the value distribution in the case where $p = p' = 0.3$. Each histogram is based on 10000 independently generated values of $\Delta^{2,3}(T^7(p), T^7(p'))$.

3. Conclusions

In each of the Tables 1–6 the minimum of every row is achieved at the diagonal element of the table. We conclude that 23-gram distance effectively discriminates between different values of the parameter.

It is interesting to study possible asymptotic distributions of the random variables $\Delta^{p,q}(T^k(p), T^k(p'))$ as well as of the corresponding subtree count vectors N^*. Empirical evidence based on a small simulation study (Example 7) suggests that $\Delta^{p,q}(T^k(p), T^k(p'))$ is asymptotically normal, for $p \neq p'$, and it is distributed as the absolute value of a normal random variable, for $p = p'$. This is not surprising as one would expect that the number N of subtrees of a bounded size would obey the central limit theorem.

References

1. N. Augsten, M. Böhlen, J. Gamper. Approximate matchiong of hierarchical data using *pq*-grams. In: K. Böhm, C.S. Jensen, L.M. Haas, M.L. Kersten, P. Larson, B.C. Ooi (Eds.), *Proceedings of the 31st International Conference on Very Large Data Bases*, Trondheim, Norway, August 30–September 2, 2005. ACM 2005, ISBN 1-59593-154-6,1-59593-177-5

REZIUMĖ

M. Bloznelis, I. Radavičius. Empirinis mažų pomedžių skaičiais paremto medžių panašumo mato tyrimas

Darbe tiriamas žymėtų medžių panašumo matas, paremtas (nedidelių) pografių skaičių palyginimu, žr. [1]. Kompiuterinio modeliavimo rezultatai rodo, kad toks panašumo matas efektyviai atskiria atsitiktinius Galtono–Vatsono medžius.

Raktiniai žodžiai: medžių palyginimas, Galtono–Vatsono medis.