CS 5433 - Big Data Management
Spring 2018
Big Data Group Project – Phase III

Group: 3

Baath, Harpinder

Polavarapu,Tejaswanth

Shankara, Akshay

Dandamudi, Madhu Kiran

Problem Understanding:

Feature selection is an important aspect of predictive analytics. It not only tends to reduce the complexity of an algorithm but improves performance and cost of data collection there by selecting sensitive variables which can provide significant predictive accuracy. The optimum feature subset selection is a NP hard problem which cannot be solved in polynomial time, genetic algorithms provide near optimal solution to this problem. Genetic algorithms are motivated from Darwin's theory of survival of the fittest and mimic biological reproduction process. It starts with random population of individuals (represented as chromosomes) and then evaluate the fitness of each individual. Two individuals are selected on the basis of their fitness value to serve as parents for crossover intending to produce better offsprings. Individuals with higher fitness value are more likely to be selected. Genetic algorithms need a lot of computational power and resources, fortunately, hadoop MapReduce allow parallelism which can be used to run genetic algorithms in parallel and obtain the desired result with improved efficiency.

In feature selection problem, each feature represents a gene and the collection of genes is a chromosome (subset of features). Each chromosome is represented by a string of 0s and 1s. 0; if feature is absent, otherwise 1.

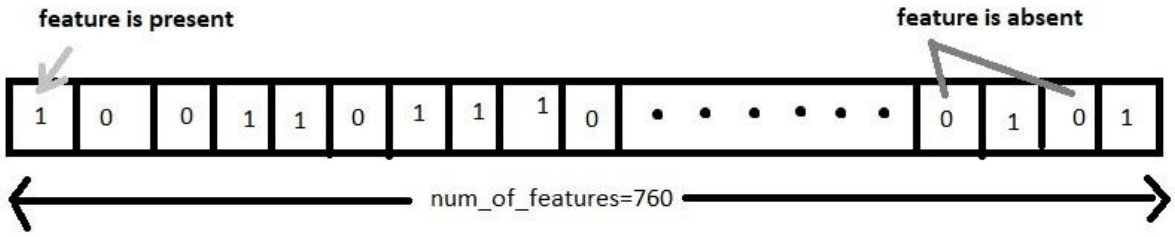

Figure 1: Representation of an individual

The population is a set of randomly selected chromosomes. With the help of the training and the test dataset obtained from the HDFS, the 'accuracy' (fitness) of each individual is calculated using machine learning algorithms such as decision tree, support vector machine etc. For each reproduction, the parents with the best fitness are selected from the pool of population using 'roulette wheel' phenomenon and a certain crossover criteria is used to produce new subsets which include features from both the parent subsets and some error (mutation). With each generation of population, we aim to achieve a better fitness and for this, we select a better set of individuals.

Implementation of GA in MapReduce

Genetic algorithm for feature selection problem would work in following steps:

1. Initialize the population with random individuals (subset of features).
2. Evaluate the fitness value of the individuals using MLib classification algorithms.
3. Select good solutions by using Tournament selection criteria (selection without replacement).

4. Create new individuals by recombining the selected population using uniform crossover.
5. Evaluate the fitness value of all offspring.
6. Repeat steps 3–5 until the convergence criteria is met.

Map()

Evaluation of the fitness function for the population (Steps 2 and 5) matches the MAP function, which has to be computed independent of other instances. As shown in the algorithm in Algorithm 1, the MAP evaluates the fitness of the given individual. It also keeps track of the the best individual and finally, writes it to a global file in the Distributed File System (HDFS). The client, which has initiated the job, reads these values from all the mappers at the end of the MapReduce and checks if the convergence criteria has been satisfied.

Algorithm 1: Map phase of each iteration of GA
map(key, value):

```
individual ← ENCODED_INDIVIDUAL(key)
fitness ← CALCULATE_FITNESS(individual)
EMIT (individual, fitness)
//Keep track of the current best

if fitness > max_fitness then
 max_fitness ← fitness
 max_individual ← individual
end if

if all individuals have been processed then
 Write best individual to global file in HDFS
end if
```

Reduce()

We will be using Tournament selection without replacement. A tournament is conducted among S randomly chosen individuals and the winner is selected. The reduce function goes through the individuals sequentially, the individuals from the last round are buffered first.When the tournament window is full, selection and crossover is carried out as shown in the Algorithm 3. When the crossover window is full, we would use the Uniform Crossover operator.The value of S would befinalized by trying different values and evaluating the result.

Algorithm 3: Reduce phase of each iteration of the GA

```
Initialize processed_individuals ← 0,
tournArray [2· tSize], crossArray [cSize]
reduce(key, values):

while values.hasNext() do
 individual ← ENCODED_INDIVIDUAL(key)
```

```
 fitness ← values.getValue()
if processed_individuals < tSize then

  //Wait for individuals to join in the tournament and   put them for the last rounds
  tournArray [tSize + processed_individuals % tSize] ← individual
 else

  //Conduct tournament over past window
  Selection_Crossover()
 end if
 processed_individuals ← processed_individuals + 1
if all individuals have been processed then

  //Cleanup for the last tournament windows
  for k=1 to tSize do

   Selection_Crossover()
   processed_individuals ← processed_individuals+1
  end for
 end if
end while

Selection_Crossover:
crossArray[processed_individuals % cSize] ← TOURN(tournArray)
if (processed_individuals - tSize) % cSize = cSize - 1 then
 newIndividuals ← CROSSOVER(crossArray)
 for individual in newIndividuals do
  EMIT (individual, dummyFitness)
 end for
end if
```

Results

Results obtained from Decision Tree and Logistic Regression are different. In addition, the number of iterations each algorithm took to get best individual are also different, that can observed in figures below.

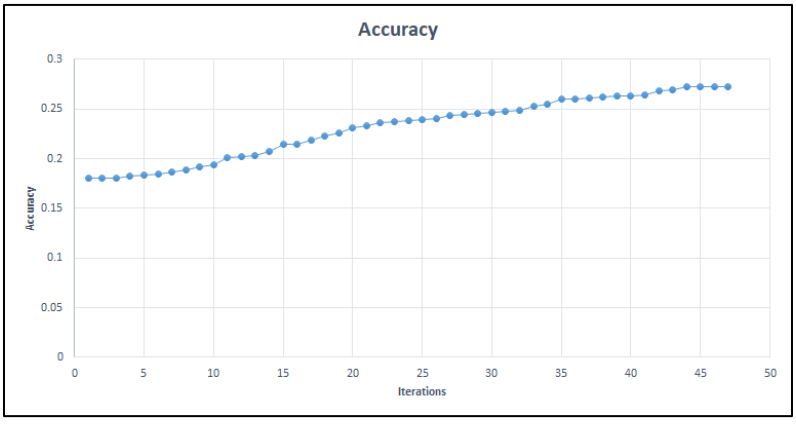

Figure 3: Effect of iterations on the Accuracy for logistic regression

Decision tree with 376 features got the accuracy of 0.272 and logistic regression with 412 features got the accuracy of 0.226.

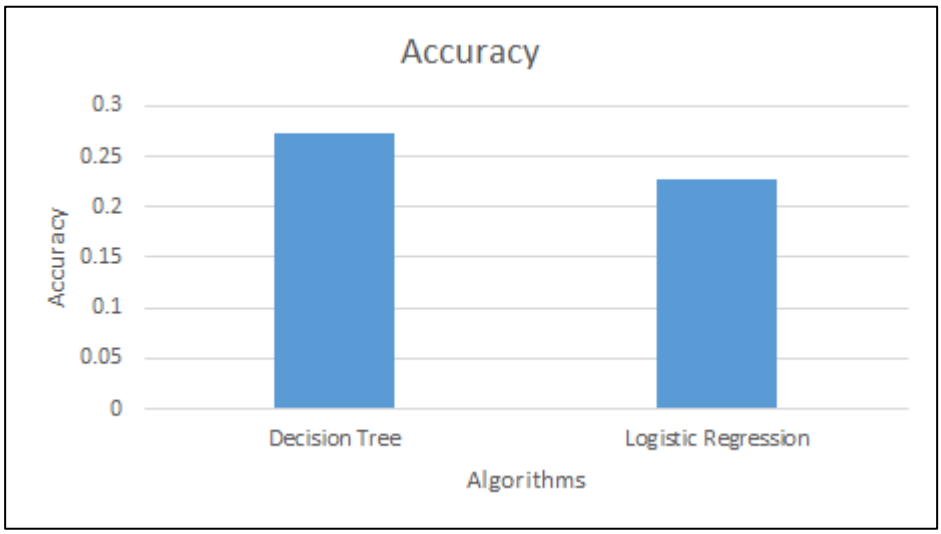

Figure 4: Prediction accuracy for optimal subset by two prediction algorithms

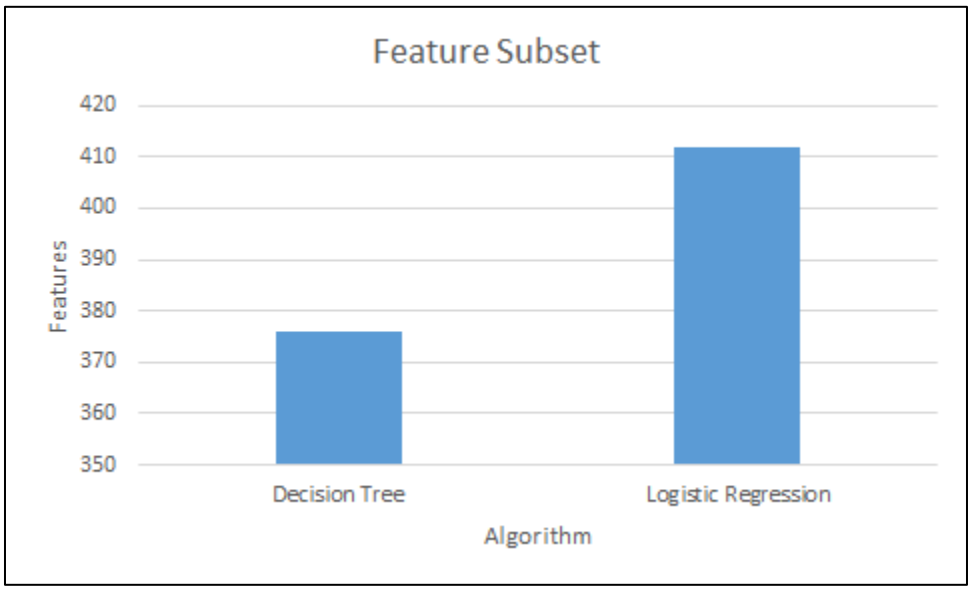

Figure 5: Size of optimal feature subset for two prediction algorithms

The optimal feature subset that we obtained from GA using decision tree algorithm is of size 376. Figure 6 shows the features in the subset. The subset obtained using logistic regression is of size 412, features included are shown in figure 7.

Feature 0	Feature 97	Feature 187	Feature 287	Feature 404	Feature 504	Feature 615	Feature 724
Feature 2	Feature 101	Feature 188	Feature 293	Feature 405	Feature 508	Feature 617	Feature 725
Feature 3	Feature 102	Feature 190	Feature 298	Feature 407	Feature 513	Feature 619	Feature 726
Feature 4	Feature 103	Feature 191	Feature 300	Feature 408	Feature 520	Feature 627	Feature 728
Feature 6	Feature 104	Feature 192	Feature 303	Feature 409	Feature 521	Feature 628	Feature 729
Feature 7	Feature 105	Feature 193	Feature 304	Feature 412	Feature 522	Feature 629	Feature 731
Feature 10	Feature 107	Feature 194	Feature 306	Feature 413	Feature 523	Feature 634	Feature 734
Feature 11	Feature 109	Feature 195	Feature 309	Feature 414	Feature 524	Feature 635	Feature 735
Feature 12	Feature 110	Feature 198	Feature 313	Feature 415	Feature 526	Feature 636	Feature 736
Feature 13	Feature 112	Feature 199	Feature 316	Feature 416	Feature 531	Feature 639	Feature 737
Feature 15	Feature 117	Feature 201	Feature 317	Feature 419	Feature 532	Feature 640	Feature 738
Feature 16	Feature 120	Feature 202	Feature 319	Feature 425	Feature 533	Feature 643	Feature 739
Feature 20	Feature 122	Feature 203	Feature 322	Feature 427	Feature 534	Feature 644	Feature 740
Feature 21	Feature 127	Feature 205	Feature 323	Feature 428	Feature 535	Feature 645	Feature 741
Feature 22	Feature 129	Feature 209	Feature 324	Feature 429	Feature 536	Feature 646	Feature 743
Feature 23	Feature 130	Feature 213	Feature 325	Feature 430	Feature 539	Feature 648	Feature 745
Feature 25	Feature 133	Feature 214	Feature 326	Feature 432	Feature 542	Feature 650	Feature 746
Feature 26	Feature 134	Feature 215	Feature 330	Feature 436	Feature 545	Feature 656	Feature 748
Feature 27	Feature 135	Feature 219	Feature 339	Feature 437	Feature 547	Feature 657	Feature 749
Feature 31	Feature 136	Feature 222	Feature 341	Feature 441	Feature 549	Feature 659	Feature 750
Feature 33	Feature 137	Feature 223	Feature 343	Feature 442	Feature 551	Feature 664	Feature 754
Feature 38	Feature 139	Feature 229	Feature 348	Feature 444	Feature 552	Feature 666	Feature 755
Feature 39	Feature 141	Feature 235	Feature 349	Feature 445	Feature 554	Feature 673	Feature 756
Feature 41	Feature 143	Feature 236	Feature 350	Feature 447	Feature 559	Feature 678	Feature 757
Feature 44	Feature 144	Feature 240	Feature 354	Feature 449	Feature 560	Feature 679	Feature 758
Feature 48	Feature 145	Feature 241	Feature 355	Feature 450	Feature 561	Feature 680	Feature 759
Feature 51	Feature 146	Feature 242	Feature 357	Feature 453	Feature 562	Feature 681	
Feature 53	Feature 147	Feature 243	Feature 358	Feature 455	Feature 564	Feature 682	
Feature 55	Feature 148	Feature 244	Feature 360	Feature 457	Feature 570	Feature 683	
Feature 56	Feature 149	Feature 245	Feature 362	Feature 459	Feature 571	Feature 688	
Feature 57	Feature 155	Feature 247	Feature 365	Feature 463	Feature 573	Feature 691	
Feature 60	Feature 156	Feature 248	Feature 367	Feature 464	Feature 574	Feature 692	
Feature 61	Feature 157	Feature 249	Feature 369	Feature 465	Feature 575	Feature 694	
Feature 62	Feature 160	Feature 254	Feature 370	Feature 466	Feature 577	Feature 695	
Feature 64	Feature 161	Feature 256	Feature 375	Feature 468	Feature 578	Feature 696	
Feature 65	Feature 162	Feature 257	Feature 379	Feature 469	Feature 580	Feature 697	
Feature 67	Feature 163	Feature 259	Feature 385	Feature 470	Feature 582	Feature 699	
Feature 69	Feature 164	Feature 260	Feature 386	Feature 472	Feature 583	Feature 700	
Feature 72	Feature 166	Feature 261	Feature 389	Feature 477	Feature 589	Feature 703	
Feature 73	Feature 167	Feature 263	Feature 390	Feature 479	Feature 590	Feature 705	
Feature 74	Feature 169	Feature 268	Feature 391	Feature 483	Feature 591	Feature 706	
Feature 78	Feature 170	Feature 269	Feature 392	Feature 488	Feature 596	Feature 707	
Feature 80	Feature 171	Feature 270	Feature 393	Feature 494	Feature 598	Feature 708	
Feature 82	Feature 172	Feature 272	Feature 394	Feature 495	Feature 599	Feature 709	
Feature 84	Feature 176	Feature 274	Feature 397	Feature 497	Feature 600	Feature 711	
Feature 85	Feature 177	Feature 275	Feature 398	Feature 498	Feature 605	Feature 712	
Feature 87	Feature 178	Feature 277	Feature 399	Feature 499	Feature 606	Feature 714	
Feature 89	Feature 181	Feature 281	Feature 401	Feature 500	Feature 609	Feature 719	
Feature 94	Feature 182	Feature 283	Feature 402	Feature 502	Feature 610	Feature 721	
Feature 96	Feature 185	Feature 285	Feature 403	Feature 503	Feature 611	Feature 722	

Figure 6: Feature subset obtained using decision tree

Feature 0	Feature 84	Feature 172	Feature 268	Feature 367	Feature450	Feature 537	Feature 641	Feature 739
Feature 1	Feature 85	Feature 176	Feature 269	Feature 368	Feature 451	Feature 540	Feature 644	Feature 740
Feature 4	Feature 87	Feature 177	Feature 270	Feature 369	Feature 454	Feature 541	Feature 645	Feature 741
Feature 6	Feature 89	Feature 178	Feature 272	Feature 370	Feature 456	Feature 543	Feature 646	Feature 742
Feature 7	Feature 92	Feature 181	Feature 274	Feature 372	Feature 458	Feature 544	Feature 647	Feature 744
Feature 10	Feature 94	Feature 182	Feature 275	Feature 375	Feature 460	Feature 546	Feature 649	Feature 746
Feature 11	Feature 96	Feature 185	Feature 277	Feature 376	Feature 464	Feature 547	Feature 651	Feature 747
Feature 12	Feature 97	Feature 187	Feature 279	Feature 377	Feature 465	Feature 548	Feature 657	Feature 749
Feature 13	Feature 101	Feature 188	Feature 281	Feature 378	Feature 466	Feature 549	Feature 658	Feature 750
Feature 15	Feature 102	Feature 190	Feature 283	Feature 380	Feature 467	Feature 550	Feature 660	Feature 751
Feature 16	Feature 103	Feature 191	Feature 285	Feature 383	Feature 469	Feature 552	Feature 665	Feature 755
Feature 20	Feature 104	Feature 192	Feature 287	Feature 386	Feature 470	Feature 553	Feature 667	Feature 756
Feature 21	Feature 105	Feature 193	Feature 293	Feature 387	Feature 471	Feature 555	Feature 674	Feature 758
Feature 22	Feature 107	Feature 194	Feature 298	Feature 390	Feature 473	Feature 560	Feature 679	
Feature 23	Feature 109	Feature 195	Feature 300	Feature 391	Feature 478	Feature 561	Feature 680	
Feature 24	Feature 110	Feature 198	Feature 303	Feature 392	Feature 480	Feature 562	Feature 681	
Feature 25	Feature 112	Feature 199	Feature 304	Feature 393	Feature 484	Feature 563	Feature 682	
Feature 26	Feature 117	Feature 201	Feature 306	Feature 394	Feature 485	Feature 565	Feature 683	
Feature 27	Feature 120	Feature 202	Feature 309	Feature 395	Feature 486	Feature 571	Feature 684	
Feature 31	Feature 122	Feature 203	Feature 313	Feature 398	Feature 487	Feature 572	Feature 689	
Feature 32	Feature 127	Feature 205	Feature 316	Feature 399	Feature 489	Feature 574	Feature 692	
Feature 33	Feature 129	Feature 209	Feature 317	Feature 400	Feature 490	Feature 575	Feature 693	
Feature 38	Feature 130	Feature 213	Feature 319	Feature 402	Feature 491	Feature 576	Feature 695	
Feature 39	Feature 133	Feature 214	Feature 322	Feature 403	Feature 492	Feature 578	Feature 696	
Feature 41	Feature 134	Feature 215	Feature 323	Feature 404	Feature 493	Feature 579	Feature 697	
Feature 44	Feature 135	Feature 219	Feature 324	Feature 405	Feature 495	Feature 581	Feature 698	
Feature 46	Feature 136	Feature 222	Feature 325	Feature 406	Feature 496	Feature 583	Feature 700	
Feature 48	Feature 137	Feature 223	Feature 326	Feature 408	Feature 498	Feature 584	Feature 701	
Feature 51	Feature 139	Feature 229	Feature 330	Feature 409	Feature 499	Feature 590	Feature 704	
Feature 53	Feature 141	Feature 235	Feature 333	Feature 410	Feature 500	Feature 591	Feature 706	
Feature 54	Feature 143	Feature 236	Feature 334	Feature 413	Feature 501	Feature 592	Feature 707	
Feature 55	Feature 144	Feature 237	Feature 335	Feature 414	Feature 503	Feature 597	Feature 708	
Feature 56	Feature 145	Feature 240	Feature 336	Feature 415	Feature 504	Feature 599	Feature 709	
Feature 57	Feature 146	Feature 241	Feature 337	Feature 416	Feature 505	Feature 600	Feature 710	
Feature 58	Feature 147	Feature 242	Feature 339	Feature 417	Feature 509	Feature 601	Feature 712	
Feature 60	Feature 148	Feature 243	Feature 341	Feature 420	Feature 514	Feature 606	Feature 713	
Feature 61	Feature 149	Feature 244	Feature 343	Feature 426	Feature 516	Feature 607	Feature 715	
Feature 62	Feature 155	Feature 245	Feature 345	Feature 428	Feature 517	Feature 610	Feature 720	
Feature 64	Feature 156	Feature 247	Feature 348	Feature 429	Feature 518	Feature 611	Feature 722	
Feature 65	Feature 157	Feature 248	Feature 349	Feature 430	Feature 521	Feature 612	Feature 723	
Feature 67	Feature 160	Feature 249	Feature 350	Feature 431	Feature 522	Feature 616	Feature 725	
Feature 69	Feature 161	Feature 251	Feature 354	Feature 433	Feature 523	Feature 618	Feature 726	
Feature 71	Feature 162	Feature 254	Feature 355	Feature 437	Feature 524	Feature 620	Feature 727	
Feature 72	Feature 163	Feature 256	Feature 357	Feature 438	Feature 525	Feature 628	Feature 729	
Feature 73	Feature 164	Feature 257	Feature 358	Feature 442	Feature 527	Feature 629	Feature 730	
Feature 74	Feature 166	Feature 259	Feature 360	Feature 443	Feature 532	Feature 630	Feature 732	
Feature 76	Feature 167	Feature 260	Feature 362	Feature 445	Feature 533	Feature 635	Feature 735	
Feature 78	Feature 169	Feature 261	Feature 363	Feature 446	Feature 534	Feature 636	Feature 736	
Feature 80	Feature 170	Feature 263	Feature 364	Feature 448	Feature 535	Feature 637	Feature 737	
Feature 82	Feature 171	Feature 265	Feature 365	Feature 449	Feature 536	Feature 640	Feature 738	

Figure 7: Feature subset obtained using logistic regression