Partial Context Similarity of Gene/Proteins in Leukemia Using Context Rank Based Hierarchical Clustering Algorithm

Shahana Bano1, K. Rajasekara Rao2
1Department of Computer Science and Engineering, K L University
2Department of Computer Science and Engineering, Sri Prakash College of Engineering

ABSTRACT

In this paper we proposed a method which avoids the choice of natural language processing tools such as pos taggers and parsers reduce the processing overhead. Moreover, we suggest a structure to immediately create a large-scale corpus annotated along with disease names, which can be applied to train our probabilistic model. In this proposed work context rank based hierarchical clustering method is applied on different datasets namely colon, Leukemia, MLL medical diseases. Optimal rule filtering algorithm is applied on these datasets to remove unwanted special characters for gene/protein identification. Finally, experimental results show that proposed method outperformed existing methods in terms of time and clusters space.

Copyright © 2015 Institute of Advanced Engineering and Science. All rights reserved.

1. INTRODUCTION

Life science studies are characterized by the construction of large and heterogeneous patterns of biological study, including protein or gene series. Therefore, a number of methods based upon text-mining have been used to improve the identify protein and genes names in medical texts. Text mining has been defined as the discovery by computer of recent, previously unknown, data by automatically extracting data from different written resources. Machine learning means the development and study of systems that could learn from data. This is actually a technique of teaching computers in order to make and enhance behaviors based on some data. Machine learning is a huge field with hundreds of algorithms for addressing different issues. Machine learning provides challenging problems in terms of algorithmic approach, data representation, computational effectiveness, and quality of the resulting program. Biomedical data along with its updates are saved in natural language style. Due to the enhanced amount of biomedical sources, it is becoming more and more challenging to find useful and relevant information regarding a specific topic. All research inventions come and enter the repository at high-rate, making the strategy of finding out and disseminating quality information a very difficult task. Manual assessment of such large amount of data will probably be very difficult and time-consuming. The issue is further magnified by the consumption of large evaluation measures, and datasets that contain essentially different annotation formats and task definitions.

Medical text documents continuously hide valuable structured data. For example, a collection of newspaper content will contain details on the location of the head-quarters of various entities. If we need to find the position of the head-quarters of, say Microsoft we could try and utilize conventional data retrieval techniques for discovering documents that contain the answer on the present query. An application of systems biology is to uncover the bio-processes underlying the patterns of a cell. Relationships within genes encode most of this data and are occasionally discovered and symbolized as key products. Understanding these
relationships is an extremely challenging issue as even the simplest organisms contain variety genes that interact in complex combinations to deal with ecological circumstances. Another complicating element is current high throughput technique designed to determine the activity level of genes is extremely noisy [8]. As there exists very few well understood genetic activities, unsupervised clustering is a common first step to understand these data.

The clustering procedure is a basic tool to organize a collection of objects within a metric space into a set of smaller partitions called clusters. By using clusters, the representation of the object pool can be made easier and the computation expense of data management can be reduced. The created clusters can be used to introduce rules of top levels describing the common characteristics of data objects. In the case of grammar induction structures, the rules of grammar are stated on word classifications as the words within the same category are transformed similarly. If word categories are known, grammar principles might be explored in a better way.

Nearest neighbor is a machine learning method introduced in the literature that often learns by comparing each individual new case to prior examples. Machine learning is definitely an area of artificial intelligence focusing on the development of approaches which permit computers to learn. More clearly, machine learning is a method for generating computer programs for the evaluation of datasets. Instance based learning, of which nearest neighbor is a subset, is a branch of machine learning techniques; other branches include: rule based genetic algorithms, ANN and support-vector-machines.

In the whole nearest neighbor algorithm, all tuples are generally saved in memory during data training. When a new query instance is accepted the memory is searched to find the instance that suits the query instance most closely. Nearest neighbor will then infer that the concept label of the query instance is similar as the notion label of the most similar instance stored in memory.

Noise present in data is a significant challenge avoiding machine learners away from being more quality, or applicable to the large selection of domains. Noise is an incorrect attribute or model value information which can be a effect of errors in manual data entry, compilation, measurement or corruption of data. If the potential for noise is certainly not recognized, this can lead to machine learning algorithms fitting the noise. Fitting the noise happens when the machine learner learns the noisy data as if were not noisy information. Noise will often make instances in memory oppose one another.

2. RESEARCH METHOD

Following are the limitations of the related work discussed in this section.

Eliminate the Non-Functional Characters
- Apply Heuristic Policies to Remove Non-Functional Symbols
- Remove and replace the following symbols with gaps: #â€œ? $â€œ®@~!\n- Remove the subsequent characters if they are followed by a space: ::
- Eliminate the following pairs of brackets if the open bracket is preceded by a space and the closed bracket is followed by a space: [ ]()
- Eliminate the single quotation symbol if it is associated with by a space or if it is preceded by a space.
- Remove s and t if they are associated with by a space.
- Eliminate slash / if it is associated with by a space.

Our proposed work overcomes all these limitations. We take three biomedical disease datasets offline to extract hidden patterns using feature extraction and hierarchical clustering approaches. Each dataset is preprocessed to remove non-functional characters to identify disease names by using gene/protein database. Hierarchical methods for supervised and unsupervised datamining give multilevel indexing of data. It can be relevant for several applications associated to data extraction, patterns retrieval and data organization.
Hierarchical Clustering Algorithm:

**Input**: Name entity Gene/Protein tags Tgp using NER approach, Gene/Protein DB, Probability P, Classes Positive pos, Negative neg, Tokenset Tk, Sentenceset Sen.

Read k, Threshold, Entropy weight;

**Output**: Quality k-abstracts.

Tgp=Get(Name _Entity_Gene/Protein_Tags)
for each tg in Tgp
For each in Tk
Calculate tag probability
List.add(tg)

---

Partial Context Similarity of Gene/Proteins in Leukemia Using Context Rank Based … (Shahana Bano)
List.add()
count=count+1
end
end.
For each token t in Tk
For each sen in Sentenceset
If((t ∈ Sen)&( t ∈ Tgp )&& ( >getProb(t))
List Data ← Sentence_id,token, Pmid, Entropy_weight,Synonyms,Data,Title,PositiveClass
Else
List Data ← Sentence_id,token, Pmid, Entropy_weight,Synonyms,Data,Title,NegativeClass
End
End
For each pair of objects in Data
Calculate distance between two objects as
\[
D(c_1, c_2) = (1 - r_{ij}) \times 0.5
\]
\[
r_{ij} = \frac{\sum_{i=1}^{d} (c_{1,i} - \bar{c}_1)(c_{2,i} - \bar{c}_2))}{\sqrt{\sum_{i=1}^{d} (c_{1,i} - \bar{c}_1)^2 \sum_{i=1}^{d} (c_{2,i} - \bar{c}_2)^2}}
\]
6. a. Start with the disjoint clustering that have level as 0 and sequence_number m = 0.
b. Rank the pairs from smallest distance (similarities in common) to the maximal distance.
c. Calculate and count pairs, say n pairs.
   If n > 0
      do,
      c.1 Explore the median as root hierarchical node.
      c.2 Split the pairs as left and right side branches based on the median.
      c.3 Explore the smallest unlike pair of clusters in the leftside and rightside current clustering, say pair rs, ls
      according to d[(rs),(ls)] = min r[(i),(j)] in which the minimum value is taken over all pairs of clusters in the
      current clustering.
      c.4 If leftside and rightside have atleast one similar object. In this case merge it collectively in one cluster,
      and look up smallest value over all pairs of clusters in the currentclustering.
      Else
      c.5 Find the maximal dissimilar pair of clusters in the leftside and rightside current clustering, say pair rs, ls
      according to d[(rs),(ls)] = max r[(i),(j)] in which the m value is taken over all pairs of clusters in the current
      clustering.
      d. Increment the sequence number: m = m + 1. (In both left and right sides) Merge clusters (r) and (s) into a
      single-cluster to form the subsequent cluster m. Place the level of this cluster to L(m) = r[(r),(s)]
      e. Revise the tree, T, by eliminating the nodes corresponding to clusters (p) and (q) and adding a node
      corresponding to the newly composed cluster. The neighborhood betweenthe new cluster, denoted (p,q) and
      old cluster (m) is stated in this way:
      d[(m), (p,q)] = min r[(m),(p)], d([(m),(q)].
      If d < 0
      Then

Minimum Variance:

The distance between two clusters is defined as the increase in the sum of squared errors (SSE) when the two clusters are merged. The SSE for a given cluster \( C_i \) is given as:

\[
SSE_i = \sum_{x \in C_i} (x - \mu_i)^2
\]

and the SSE for a clustering \( C = \{C_1, \ldots, C_m\} \) is given as:

\[
SSE = \sum_{i=1}^{m} SSE_i
\]

When we merge \( C_i \) and \( C_j \), the change in the SSE involves these three clusters, and is given as:

\[
\text{ASSE}_{ij} = SSE_{ij} - SSE_{i} - SSE_{j}
\]

Plugging into the equation above, after simplification, we thus obtain the distance between the two clusters as:

\[
d(C_i, C_j) = \Delta SSE_{ij} = \frac{\sum_{x \in C_i} \sum_{y \in C_j} (x - \mu_i - y + \mu_j)^2}{\sum_{x \in C_i} \sum_{y \in C_j} (x - \mu_i - y + \mu_j)^2}
\]
Partial Context Similarity of Gene/Proteins in leukemia:

Context Similarity %5.3f====>0.2526455026455026

U19107_at synoym are ZNF127 (ZNF127) gene
Context Similarity %5.3f====>0.3436507936507936

U19142_at synoym are GAGE1 G antigen 1 (GAGE-1)
Context Similarity %5.3f====>0.4829059829059829

U19180_at synoym are BAGE B melanoma antigen
Context Similarity %5.3f====>0.4363929146537842

U19261_at synoym are Epstein-Barr virus-induced protein mRNA
Context Similarity %5.3f====>0.2578347578347578

U19345_at synoym are AR1 protein (AR) mRNA
Context Similarity %5.3f====>0.4391534391534391

U19487_at synoym are Prostaglandin E2 receptor mRNA
Context Similarity %5.3f====>0.2629629629629629

U19517_at synoym are (apoargC) long mRNA
Context Similarity %5.3f====>0.38791423001949316

U19523_at synoym are GCH1 GTP cyclohydrolase 1 (dopa-responsive dystonia) [alternative products]
Context Similarity %5.3f====>0.41629629629629633

U19718_at synoym are MFAP2 Microfibrillar-associated protein 2
Context Similarity %5.3f====>0.3785004516711834

U19796_at synoym are Melanoma antigen p15 mRNA
Context Similarity %5.3f====>0.43407407407407406

Partial Context Similarity of Gene/Proteins in Leukemia Using Context Rank Based ... (Shahana Bano)
<=== U19878_at ===> synonyms are Transmembrane protein mRNA
   Context Similarity %5.3f ===> 0.43304843304843305

<=== U19906_at ===> synonyms are VASOPRESSIN V1A RECEPTOR
   Context Similarity %5.3f ===> 0.0

<=== U19948_at ===> synonyms are Protein disulfide isomerase (PDIp) mRNA
   Context Similarity %5.3f ===> 0.2578347578347578

<=== U19977_at ===> synonyms are Preprocarboxypeptidase A2 (proCPA2) mRNA
   Context Similarity %5.3f ===> 0.42407407407407405

<=== U20158_at ===> synonyms are 76 kDa tyrosine phosphoprotein SLP-76 mRNA
   Context Similarity %5.3f ===> 0.42328042328042326

<=== U20230_at ===> synonyms are "GB DEF = Guanyl cyclase C gene, partial cds"
   Context Similarity %5.3f ===> 0.37777777777777777

<=== U20240_at ===> synonyms are "CEBPG CCAAT/enhancer binding protein (C/EBP), gamma"
   Context Similarity %5.3f ===> 0.4199860237596087

<=== U20285_at ===> synonyms are Gps1 (GPS1) mRNA
   Context Similarity %5.3f ===> 0.0

<=== U20325_at ===> synonyms are Cocaine and amphetamine regulated transcript CART (hCART) mRNA
   Context Similarity %5.3f ===> 0.41816009557945044

<=== U20350_at ===> synonyms are CMKRL1 Chemokine receptor-like 1
   Context Similarity %5.3f ===> 0.38078703703703703

<=== U20362_at ===> synonyms are Tg737 mRNA
   Context Similarity %5.3f ===> 0.4037037037037037

<=== U20391_rna6_at ===> synonyms are Folate receptor (FOLR1) gene
   Context Similarity %5.3f ===> 0.32936507936507936

<=== U20428_at ===> synonyms are SNC19 mRNA sequence
   Context Similarity %5.3f ===> 0.0

<=== U20530_at ===> synonyms are GB DEF = Bone phosphoprotein spp-24 precursor mRNA
   Context Similarity %5.3f ===> 0.37703703703703706

Correlation Distance Metric:
Correlation Distances: 0.5246662304900925
Correlation Distances: 0.5619362422999764
Correlation Distances: 0.651394771224407
Correlation Distances: 0.48759512587181975
Correlation Distances: 0.5319049159237761
Correlation Distances: 0.5246662304900925
Correlation Distances: 0.5619362422999764
Correlation Distances: 0.651394771224407
Correlation Distances: 0.5319049159237761
Correlation Distances: 0.5246662304900925
Correlation Distances: 0.5619362422999764
Correlation Distances: 0.651394771224407
Correlation Distances: 0.5319049159237761
Correlation Distances: 0.5619362422999764
Correlation Distances: 0.651394771224407
Correlation Distances: 0.5319049159237761
Correlation Distances: 0.6221864849879517
Correlation Distances: 0.6058234775837336

Clustered Instances
0 11 (92%)
1 1 (8%)

ACCURACY DETAILS
TOTAL GENE DETECTION ACCURACY 12 100%
ERROR RATE OF PROPOSED ALGORITHM 0 0%
Correlation Efficiency 1
Total Number of Instances 12
4. CONCLUSION

In this proposed work context rank based hierarchical clustering method is applied on different datasets namely colon, Leukemia, MLL medical diseases. Optimal rule filtering algorithm is applied on these datasets to remove unwanted special characters for gene/protein identification. This work overcomes some of the limitations in the literature such as: noise elimination in medical datasets, robustness, high disease prediction rate, high quality cluster result with less search space and high true positive rate. Finally, experimental results show that proposed method outperformed well in terms of time and clusters search space are concerned. In future this work can be extended to implement similar disease clusters on online medical documents like medline, pubmed etc.

REFERENCES


BIOGRAPHIES OF AUTHORS

Shahana Bano received her MS (IS) degree in Computer Science from Montessori Mahila Kalasala Vijayawada. M.Tech degree in Computer Science from K.L. College of Engineering Vaddeswaram and pursuing her Ph.D from KL University. Currently, she is working as a Assistant Professor in the Department of Computer Science & Engineering in KL University. She has got 7 years of teaching experience. She has published Eleven research papers in various national and international Journals. She is member of professional societies CSI.

Dr.K. Rajasekhara Rao received his Ph.D from Acharya Nagarjuna University. Currently, he is working as a Professor in the Department of Computer Science & Engineering in Sri Prakash College Of Engineering. He has got 28 years of teaching experience. He has published 30 research papers in various national and international Journals. He is member of professional societies CSI. He was awarded the "Best Dean" on 30th December 2012, organized by ASDF's.