Identifying Cancer Biomarkers Via Node Classification within a Mapreduce Framework

Taysir Hassan A. Soliman
Associate Professor, Information Systems Department, Faculty of Computers & Information, Assiut University, Egypt

***Abstract*—Big data are giving new research challenges in the life sciences domain because of their variety, volume, veracity, velocity, and value. Predicting gene biomarkers is one of the vital research issues in bioinformatics field, where microarray gene expression and network based methods can be used. These datasets suffer from the huge data voluminous, causing main memory problems. In this paper, a Random Committee Node Classifier algorithm (RCNC) is proposed for identifying cancer biomarkers, which is based on microarray gene expression data and Protein-Protein Interaction (PPI) data. Data are enriched from other public databases, such as IntACT1 and UniProt2 and Gene Ontology3 (GO). Cancer Biomarkers are identified when applied to different datasets with an accuracy rate an accuracy rate 99.16%, 99.96% precision, 99.24% recall, 99.16% F1-measure and 99.6 ROC. To speed up the performance, it is run within a MapReduce framework, where RCNC MapReduce algorithm is much faster than RCNC sequential algorithm when having large datasets.**

Keywords—Big data; cancer biomarkers; MapReduce; node classification

I. INTRODUCTION

Bioinformatics is one of the main applications that adopt big data through microarray gene expression analysis, next generation sequencing, text mining of literature publications, and large graph analysis of biological networks, such as metabolic networks, signal pathways, and protein- protein interaction networks. Bioinformatics researchers have an excellent opportunity to achieve scalable efficient and reliable computing performance on Linux clusters and within cloud computing environment [1]. However, scalable and efficient data mining algorithms are needed to perform different tasks in bioinformatics. Biomarkers play an important role in diagnosing, assessing prognosis and directing treatment of cancer. A cancer biomarker refers to a substance or process that is indicative of the presence of cancer in the body. A biomarker may be a molecule secreted by a tumor or a specific response of the body to the presence of cancer. Genetic, epigenetic, proteomic, glycomic, and imaging biomarkers can be used for cancer diagnosis, prognosis, and epidemiology4. Biologists can now quickly identify hundreds, and even thousands of candidate genes associated with a target disease or functionality. One of the main traditional techniques to find interactions and similar structure is applying text mining techniques to literature abstracts, i.e. through PubMed5 [2,3]. However, this is a very time consuming issue because of the tremendous high volume of current literature reviews.

Other techniques fall into two main categories: Mircoarray gene expression analysis and biological networks. Microarray gene expression analysis can measure thousands of gene expressions which make it a good chance to identify biomarkers through microarray technology [4-6]. However, better prediction accuracy is required since the accuracy of applying network techniques is relatively low. Identifying significant gene sets or pathways involved in diseases or biological processes by incorporating some prior biological knowledge, such as gene set enrichment analysis or pathway enrichment analysis are proposed via several methods [7-9]. In addition, PPIs, protein-DNA interactions, or regulatory pathways algorithms are developed. For instance, Chuang et al. [10] identified biomarkers of metastasis using breast cancer gene expression data, based on protein-protein interaction networks. Li et al. [11] introduced a network-constrained term based on L1-norm of regression coefficients of microarray data. Jahid and Ruan [12] identified a small number of intermediate genes containing important information about the pathways involved in metastasis genes, using a randomized steiner tree. Zhu *et al*. [13] recently built binary classifiers as prediction models, using support vector machines. In addition, Wei and Li [14] developed a Markov Random Field Model for network-based Analysis. Furthermore, Chen et al. [15] developed network-constrained Support Vector Machine (netSVM) for cancer biomarker identification with an improved prediction performance. Hwang et al. [16] applied the network propagation algorithm to study three large-scale breast cancer datasets, achieving competitive classification performance. Xia et al [17] have developed Network Analyst, enabling high performance network analysis with rich user experience in order to identify genes/ proteins of interest in biological networks.

One of the main computational challenges have become increasingly important is using High Performance Computing (HPC) in bioinformatics data analysis [18]. Another computer architecture / service model is cloud computing [19-21], where it is used to scale up the performance of the required service. Recently, biomarker prediction based on large-scale feature selection and MapReduce has been discussed in [22], where Kmeans clustering and Signal to Noise Ratio have been combined with optimization technique as Binary Particle Swarm Optimization. A key problem arises when using hybrid approaches of microarray gene expression and network-based methods is handling very large networks which require high performance time.

[1]http://www.ebi.ac.uk/intact/
[2]http://www.uniprot.org/, [3]http://geneontology.org/ ,
[4] https://en.wikipedia.org[5] http://www.ncbi.nlm.nih.gov/

In this paper, a node classification algorithm is suggested in order to identify biomarkers, which is considered one of the main problems in the bioinformatics domain. This algorithm is applied and compared to other machine learning algorithms, such as naïve bayes and random forest. In addition, the RCNC algorithm is applied within MapReduce framework, as one of the open source Apache Hadoop project. Node classification has been previously introduced in dynamic content-based networks [23]. The main contributions of this paper are:

1) A hybrid approach of microarray gene expression and PPI networks is proposed to predict protein biomarkers via Random Committee Node Classifier algorithm (RCNC).

2) Speeding up the performance of the algorithm via MapReduce.

3) Developing an information topological PPI network

The organization of this paper as follows: section two explains materials and methods and section three illustrates results and discussion. Finally, section four concludes the work and gives insights into future work.

II. Materials and Methods

In this section, identifying biomarkers based on node classification within a MapReduce framework is proposed, as illustrated in Fig. 1. This framework depends on a hybrid approach of microarray gene expression data and PPI network. The framework consists of two main phases: data preprocessing and biomarker identification, which will be discussed in details in the following subsections. Data preprocessing phase has two main goals, which are 1) Computing Differentially Expressed Genes (DEGs) and 2) Integrating data. The goal of biomarker identification phase is to identify biomarkers for different types of cancer (Breast, colon, ovarian and hepatocellular carcinoma), using the proposed RCNC algorithm.

A. Phase i: data preprocessing

The objectives of this phase are to a) Compute Differentially Expressed Genes (DEGs) and b) Integrate Data.

1) Computing deg:

Microarray technologies now enable the simultaneous interrogation of the expression level of thousands of genes to obtain a quantitative assessment of their differential activity in a given tissue or cell. Microarray analysis has enabled the identification of gene signatures for diagnosis, molecular characterization, prognosis and treatment prediction. Microarray gene expressions data are obtained from GEO[4] database for Breast, colon, liver (hepatocellular carcinoma), and ovarian cancer. For each type of cancer, five series are used, which are illustrated in Table I, where both healthy and unhealthy microarray gene expression series are downloaded (Affymetrix experiments). Differentially Expressed Genes (DEGs) are computed for all downloaded samples, using R statistical language [4]; in addition, p value < 0.05 is set as the threshold for DEGs and t-test [23] is applied.

2) Integrating data

Data integration is one of the vital tasks in bioinformatics, where many diverse public databases' formats exist, such as

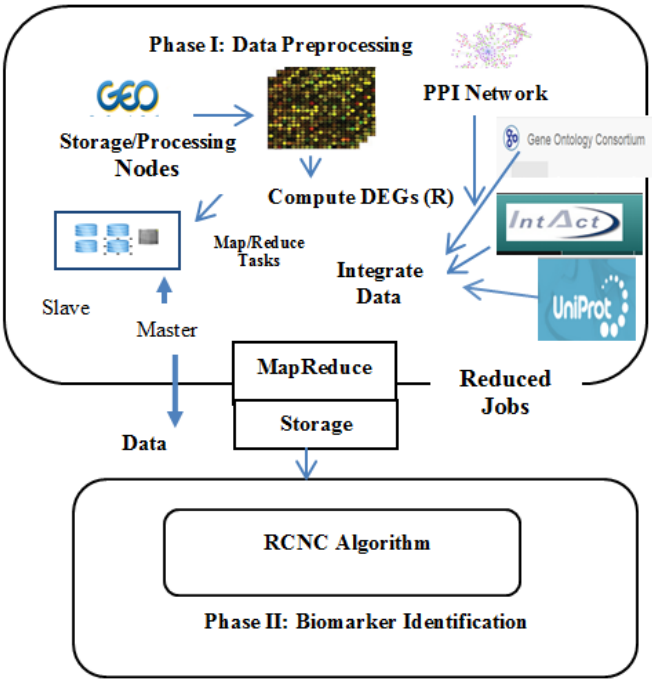

Fig. 1. Biomarker Identification Framework using RCNC Algorithm

TABLE I. GEO Cancer Series

Cancer Type	Series (Samples)	# of Gene Instances
Breast	GSE44024 (4)	22,278
	GSE53394 (8)	22,278
	GSE38376 (18)	48,804
	GSE45804 (12)	33,298
	GSE41816 (36)	33,298
Liver (Hepatocellular carcinoma)	GSE41804 (40)	54,676
	GSE49515 (26)	54,676
	GSE21955 (22)	24,527
	GSE29084 (4)	54,676
	GSE32474 (174)	54,676
Ovarian	GSE31432 (23)	48,804
	GSE51373 (28)	54,676
	GSE22600 (15)	54,676
	GSE23616 (15)	20,603
	GSE13525 (12)	54,589
Colon	GSE14773 (4)	54,676
	GSE34299 (4)	
	GSE18088 (53)	
	GSE18560 (12)	

XML, csv, and RDF. PPI data sometimes are not enough to identify biomarkers. As a result, in this approach data are integrated from heterogeneous resources: IntAct (release 2.5) and UniProt (August 2015) in addition to the DEGS results of micorrary gene expressions, computed at step 2.1.a.

In this work, cancer interaction datasets are downloaded from IntAct, which contain the target types of cancer discussed here: breast cancer, ovarian cancer, hepatocellular carcinoma, and colon cancer. The following preprocessing steps are accomplished for IntAct and UniProt data:

1) *Removing missing values*
2) *Deleting irrelevant attributes*
3) *Extracting data*
4) *Mapping attributes*

To illustrate the idea, downloaded cancer interaction data contain UniProtkb identifiers of interacting proteins, alternative identifiers for each protein at IntAct database European Bioinformatics Institute identifier, aliases, interaction detection method (two hybrid, pull down, etc), publication date of each, taxonomy identifier, interaction type (physical association, colocalization, direct interaction, and association), database source, interaction identifier, and confidence. Some of the GO ontologies are missing so the corresponding values are deleted. In addition, irrelevant attributes (attributes not used as parameters for determining biomarkers) are deleted: the publication date, taxonomy identifier, interaction detection method, interaction identifier and source database.

Gene name is extracted from attribute (Alias), and mapped to the DEGs found in microarray experiments. For example protein A: uniprotkb: P35125-3 Ubiquitin carboxyl-terminal hydrolase 6 (alternative identifier: intact:EBI-954590), interacts with protein B uniprotkb:P10916 (alternative identifier: intact:EBI-725770|uniprotkb:Q16123). In addition, alias of P35125-3 is psi-mi:p35125-3(display_long)|uniprotkb:"210(ORF1)"(isoform synonym)| uniprotkb: oncTre210p (isoform synonym)| uniprotkb: USP6(gene name)|psi-mi: USP6 (display_short)| uniprotkb:HRP1 (gene name synonym)|uniprotkb:TRE2(gene name synonym) |uniprotkb: Deubiquitinating enzyme 6(gene name synonym) |uniprotkb: Proto-oncogene TRE-2(gene name synonym)| uniprotkb: Ubiquitin-specific- processing protease 6 (gene name synonym)| uniprotkb:Ubiquitin thioesterase 6 (gene name synonym), alias of protein B is psi-mi:mlrv_human (display_long) |uniprotkb: MYL2(gene name)|psi-mi:MYL2(display_short). In addition, other attributes are interaction detection method (psi-mi:"MI:0018"(two hybrid)), publication 1st author (Dechamps et al. (2006)), publication identifier (pubmed:16555005), Taxid interactorA (taxid:9606(human)|taxid:9606(Homo sapiens), Taxid interactorB (taxid: 9606 (human) | taxid:9606(Homo sapiens)), interaction type (psi-mi:"MI:0915"(physical association)), source database(s) (psi-mi:"MI:0469"(IntAct)), interaction identifier (intact:EBI-1225898), and confidence value (intact-miscore:0.61).

For each protein, each UniProtkb identifier is mapped into its corresponding Uniprotkb identifier in UniProtkb database. Other included information from UniProtkb is protein function, Gene Ontology (GO) molecular function, biological process, and cellular component. In addition, DisGeNet database has been used as for validation of biomarkers' prediction results.

B. Phase II: Biomarker Identification

To identify biomarkers, RCNC algorithm is proposed, which depends on topological node classification algorithm in an ensemble learning manner. The problem of node classification has been addressed in a number of applications, such as social network analysis [25]. In this section, RCNC algorithm of biomarkers identification is explained in details. RCNC uses a random committee technique, which is an ensemble tree classifiers based. Ensemble methods like combine the decisions of multiple hypotheses are some of the strongest existing machine learning methods [26-28]. Ensemble classifiers gather randomizable base classifiers, where each base classifier is built using a different random number seed. A random committee algorithm is an ensemble of random tree classifiers, where it predicts a class label by averaging probability estimates over these classification trees. This algorithm produces better overall accuracy for all testing cases than any individual committee member. In this paper, a random committee technique is used to handle: 1) too large data volume, 2) inadequate data, and 3) complexity of decision boundary. The learning procedure for ensemble algorithms can be divided into the following two parts:

1) Constructing base classifiers/base models: In this part, data preprocessing is performed first where noisy data are removed then base classifier are constructed. Data preprocessing step is already at the data integration phase, as previously explained.

2) Voting: The main objective of this part is to combine the base classifiers models built in the previous step into the final ensemble model. There are several kinds of voting but the most used ones are the weighted and un-weighted voting. Voting includes the weighted average (of each base classifier holds) when using regression problem and majority voting when doing classification and the weighted-majority output is given by, which is used in this paper:

$$\text{Argmax}\left[\sum_{i=1}^{k} p_i(x), w_i\right] \quad (1)$$

Pi(x) is the results of the prediction of ith prediction model and Pi(x, w) is indicator function defined as:

$$p_i(x,w) = \begin{cases} 1 & x = w \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Problem Definition: given a graph, which is represented as G= {V, E, W}, where V is a set of nodes, E is the set of Edges, and W is the edge weight matrix n x n; W = [wij] and n = |V|. L is the set of labels L= {l1, l2, …, lq} for the set of q attributes associated with each node V.

Homophily: is a term used in social networks and defined as a link between individuals (i.e. friendship or other social connection) when they are being similar in nature. When applying "homophily" to PPI information network, two protein nodes are connected based on "homophily" property if they interact with each other and have similar characteristics. These characteristics include:

- Sequence similarity scores.
- GO relations where two nodes are GO related if there is a semantic relation holding between those proteins. This semantic relation between two proteins is divided into the following:

- If functions are connected through ontology
- If cellular components relations exist.
- If Biological process relations exist.

For example, for the protein P35125 which is a biomarker for ovarian cancer interacts with protein Q8N8A2. P35125 has gene molecular functions: calmodulin binding, cysteine-type endopeptidase activity, nucleic acid binding, ubiquitin-specific protease activity. Q8N8NA2 has a protein binding molecular function, where calmodulin binding is a protein binding type. P35125 and Q8N8A2 proteins have 84.3% sequence similarity. Sequence similarity scores are taken into consideration when >70%, as shown in Fig. 2. Table II explains the steps of graph construction algorithm.

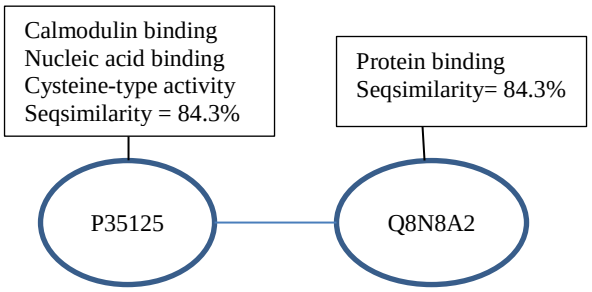

Fig. 2. An example of Breast Cancer PPI Information Network

TABLE II. GRAPH CONSTRUCTION ALGORITHM

Algorithm 1: Graph Construction

```
map(key, value):
  begin
    edge = 1;
    Node V(edge);
    If homophily exists
      Emit(V.id, V);
  end
reduce(key, values):
  begin
    Emit(key, serialize(values));
  End
```

Machine learning algorithms have the advantage of making use of Hadoop distributed computing platform and the MapReduce programming model to process data in parallel. Many machine learning algorithms have been investigated to be transformed into the MapReduce paradigm in order to make use of the Hadoop Distributed File System (HDFS). In the current work, RCNC is run under the MapReduce framework and is evaluated on four datasets in order to evaluate scalability comparisons of using RCNC sequentially and RCNC under the MapReduce environment (RCNC MapReduce). The proposed MapReduce architecture used for this classifier is clarified in Fig. 3.

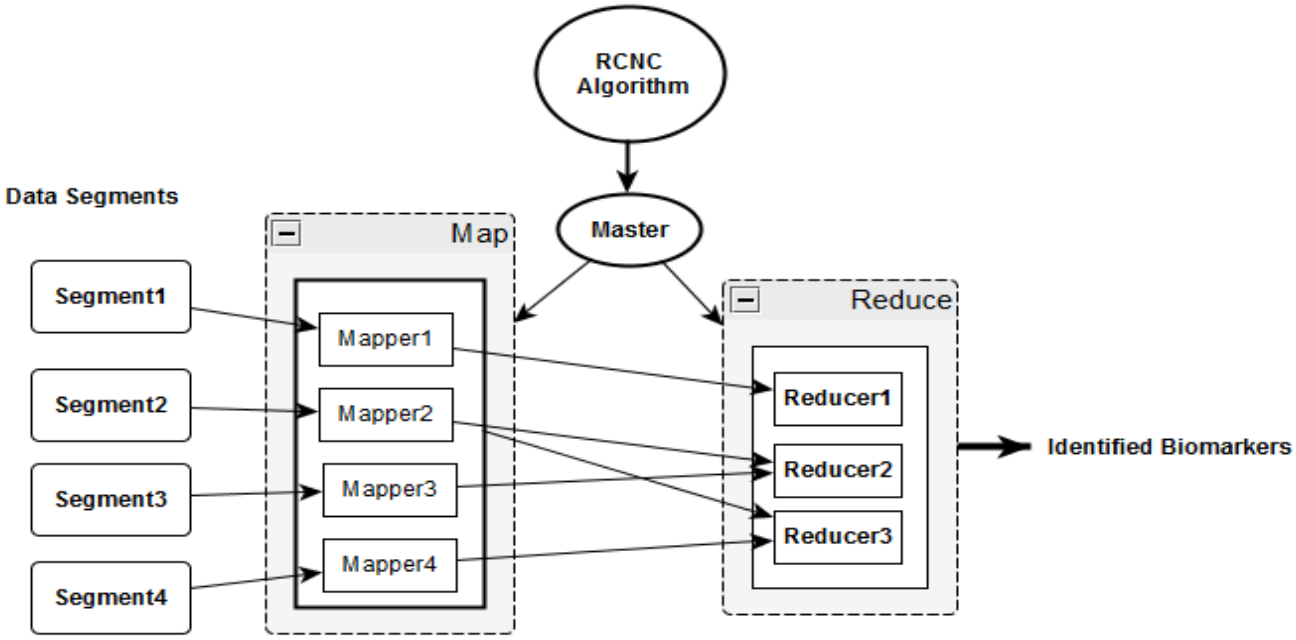

Fig. 3. Workflow of the proposed MapReduce framework for RCNC Node Classifier

Through this architecture, the number of occurrences of an attribute with a specific value given a certain class is obtained. The Hadoop uses Input Data Format to divide the big file into small input files which record Key and Value. In this case, the key will be the feature of the data (i.e. interaction type). Then, the Map process defines the data structure (key, value) on the Map operation. The Map process is applied to each input dataset in parallel. With the result from the MapReduce task, one can assign the instance to a class after training each segment via a random committee algorithm. Finally, the ensemble of classifiers is computed via equation (2). Table III illustrates the steps of RCNC MapReduce algorithm. RCNC sequential is the same idea but without dividing the algorithm into Map & Reduce functions.

TABLE III. MAPREDUCE RCNC ALGORITHM

	Algorithm 2: Random Committee Node Classifier (RCNC) MapReduce
Input	Graph G=(V, E,W), T= ensemble size; Max= Maximum number of nodes
Output	f = ($f^{(1)}$ … $f^{(T)}$) (ensemble of classifiers)

Process	Map(Vertexid V.id, Vertex V)
	Begin
	For E ε n.adjancylist do
	emit(E.neighbor, <V.label, E.EdgeWeight>)
	End
	Emit(Vertexid V.id, Vertex V)
	End
	Reduce(V.id, W)
	Begin
	For i = 1 to Max do
	Begin
	f(i) <- (p_i(V.id, W))
	End
	V.label ← ($f^{(1)}$ … $f^{(T)}$)
	Emit(Vertexid V.id, Vertex V)
	End

III. RESULTS

In this paper, four kinds of cancer are used: breast, colon, liver (hepatocellular carcinoma) and ovarian interaction datasets. Data are split into 66% for training and the rest for testing within a 10-Fold validation on the training dataset to select the optimal value of parameters. Experiments have been performed using Java JDK version 1.7 and for MapReduce implementation Hadoop version 2.4.1. MapReduce implementation is tested in a cluster of 4 data nodes running Linux. Each node is an Intel ® Core TM i7-3770 CPU @3.4 GHZ, and 32GB RAM. Several comparisons are performed: 1) the proposed RCNC algorithm for node classification in a sequential manner versus naïve bayes, random forest classifiers, proposed method in [22], and [29], as shown in Table IV. In [29], an approach based on Neighborhood Rough Set and Probabilistic Neural Networks Ensemble is proposed for the classification of Gene Expression Profiles. Comparison contains the precision, recall, F1-measure, and ROC.

As summarized in Table IV, RCNC is always higher than Random Forest and naïve bayes classifiers when for all datasets. For example, for breast cancer dataset, RCNC has shown an accuracy of 99.72% , a recall of 99.7%, ROC of 100%, where the True positive rate is 99.7% and False Positive rate is 0.05% with F1-measure 99.7% for breast cancer datasets. For ovarian datasets, both datasets 15,154 and 54,675 are tested for all algorithms: RCNC, Random Forest, naïve bayes, BSMO, and [34]. In the first case, RCNC is higher than BSPO and [34], where in the second case RCNC and BSMO give the same accuracy rate. However, RCNC gives more information regarding related biomarkers from the PPI information network. Furthermore, datasets are enlarged to 4GB each synthetically and the accuracy is the same but performance time is very fast.

The second testing of RCNC MapReduce is its time performance versus RCNC MapReduce, as illustrated in Fig. 6, where the time of RCNC MapReduce is faster than RCNC sequential.

TABLE IV. COMPARISONS OF RCNC WITH OTHER CLASSIFIERS

# Genes	Classifiers	P %	Rec %	ROC %	F1-	Acc. %
Breast 22,278	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	98.9	98.9	99.9	98.7	98.8
	Naïve Bayes	98.3	98.3	100	98.2	98.3
Colon 15,154	RCNC	96	97.4	98	97	97
	Random Forest	83	84	84	95	84.1
	Naïve Bayes	81.8	82.8	81.8	95	82
Hepato 24,527	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	80.1	38.8	83.6	88.6	75.7
	Naïve Bayes	76	76.4	76.1	81.6	75.7
Ovarian 15,154	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	81.5	79.1	90.6	81.5	81.4
	Naïve Bayes	96	97.4	98	97.1	97
	BSPO[23]					99
	[34]					96
Ovarian 54,675	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	81.5	79.1	90.6	81.5	81.4
	Naïve Bayes	96	97.4	98	97.1	97
	BSPO[23]					100
	[34]					96

Finally, Fig. 7 clarifies the runtime of RCNC MapReduce having one, two, and four nodes for each dataset. Experiments for different size of data chunk and different number of maps are performed to evaluate impact of MapReduce parallelism. One can notice that having two nodes, the time performance is reduced to near half of the time required when having one node only. In addition, having four nodes, the runtime of the algorithm is reduced. The accuracy rate of RCNC sequential versus RCNC MapReduce is also tested when having four nodes, where the accuracy remains the same.

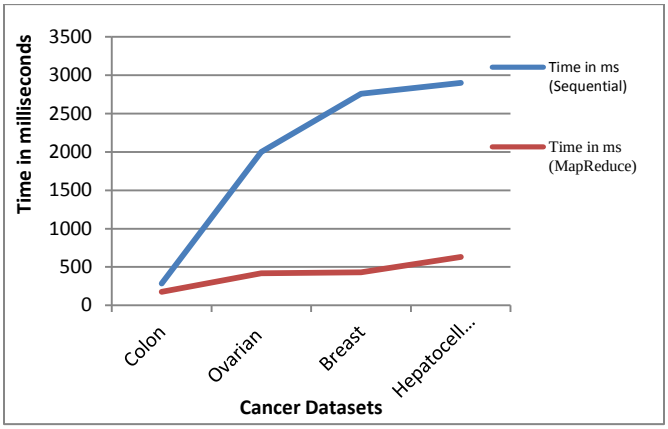

Fig. 4. Comparison of RCNC Sequential and RCNC MapReduce

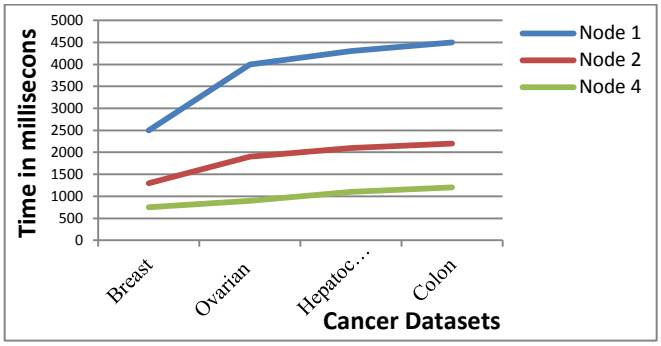

Fig. 5. Time Comparisons of RCNC MapReduce for Different Number of Nodes

Identified genes are evaluated against the DisGeNet database, where the relation between genes as biomarkers can be downloaded for cancer datasets. Examples of cancer detected biomarkers are: HSP60 (ovaries), HSPD1 (ovaries), FANCD2 (breast), FANCD3 (breast), FANCD4 (breast), MYL2 (breast), FANCD1 (ovaries), FACD (ovaries), XRCC9 (breast), DGKI (breast), APCS (colon), STK11 (colon), PTEN (colon), MLH1 (colon), MLH6 (colon), POLE (colon), EPCAM (colon), and MYH (colon)

IV. Conclusions

In this paper, a Random Committee Node Classifier algorithm (RCNC) was proposed to predict cancer biomarkers, where microarray gene expression and network based methods were used. These datasets had a very large volume, which caused main memory problems. Compared with other classifiers, RCNC had proven high accuracy. Biomarker genes were identified when applied to different datasets with an accuracy rate 99.16%, 99.96% precision, 99.24% recall, 99.16% F1-measure and 99.6 ROC. To speed up the performance, it was run within a MapReduce framework, where RCNC MapReduce were much more faster than RCNC sequential when having large datasets. Future work includes taking RNAseq data into consideration and enlarging the datasets into multiple types of cancer. In addition, more ontologies will be added as ChEBI and disease ontologies. Furthermore, more enhancements can be performed to RCNC for covering multi-dimensional graphs.

References

[1] Q. Zou, X. , Li,, W. Jiang,, Z. Lin, G. Li, and K. Chen,. "Survey of mapReduce frame operation in bioinformatics," Briefings in Bioinformatics, pp. 1-12, 2013.

[2] H. Li and C.Liu, "Biomarker identification using text mining," Computational and Mathematical Methods in Medicine, pp. 1-4, 2012..

[3] W.Fleuren, et al., "Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature minin," BioDataMining, Vol. 6, 2, pp.1-15, 2013.

[4] T. Golub, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science," Vol. 286, issue 5439, pp.531-537, 1999.

[5] C. Sotiriou, and L. Pusztai , "Gene-expression Signatures in breast cancer," The New England Journal of Medicine, 360, 8, pp. 790-800, 2009.

[6] V. Tusher, , R. Tibshirani, and G. Chu, G., "Significance analysis of microarrays applied to the ionizing radiation response," Proc. Natl Acad Sci USA, Vol. 98, 9, pp.5116-5121, 2001.

[7] T. Bo,and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," Genome Biology, Vol. 3, issue 4, pp. 1-11, 2002.

[8] A. Subramanian, et al., "Gene set enrichment analysis: a Knowledge-based Approach for interpreting genome-wide expression profiles," Proceedings of National Academy Science, 102, 43, pp.15545-15550, 2005.

[9] R., Curtis, M. Oresic, and A.,Vidal-Puig, "Pathways to the analysis of microarray data," Trends Biotechnology, 23, 8, pp. 429-435, 2005.

[10] H. Chuang, , E. Lee, Y. Liu, , D. Lee, and T. Ideker. "T: network-based classification of breast cancer metastasis," Molecular Systems Biology, 3, 140, 2007.

[11] C. Li and H.Li, "Network-constrained regularization and variable selection for analysis of genomic data," Bioinformatics, 24, 9, pp. 1175-1182, , 2008.

[12] M. Jahid and J. Ruan, "A steiner tree-based method for biomarker discovery and classification in breast cancer metastasis," BMC Genomics, 13 (Suppl 6):S8, pp. 1-9, 2012. .

[13] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," BMC Bioinformatics, 10 (Suppl 1):S21, 2009.

[14] Z. Wei and H. Li, "A markov random field model for network-based analysis of genomic data," Bioinformatics, 23, 12, 1537-1544, 2007.

[15] L.Chen, , J. Xuan, R. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," BMC Systems Biology, 5, 16, 2011.

[16] Hwang et al., "Robust and efficient identification of biomarkers by classifying features on graphs," Bioinformatics, Vol. 24, 18, pp.2023–2029, 2008.

[17] J. Xia, M. Benner, and R. Hancock, "NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration," Nucleic Acids Research. 42, 167-174, 2014.

[18] R. Taylor, "An overview of the hadoop/mapReduce/HBase framework and its current applications in bioinformatics," BMC Bioinformatics, 11(Suppl 12):S, 1-6, 2010.

[19] C. Sansom, "Up in a cloud?. Nature Biotechnology, 28, 1, pp.13-15, 2010.

[20] L. Stein, "The case for cloud computing in genome informatics," Genome Biology, 11:207, 2011.

[21] M. Schatz, B. Langmead, and S. Salzberg, "Cloud computing and the DNA data race," Nature Biotechnology, 28, pp.691-693, 2011.

[22] A. Kourid, and M. Batouche, " Biomarker discovery based on large-scale feature selection and mapreduce," Proceedings of the 5th IFIP TC 5 International Conference, CIIA 2015, Saida, Algeria, May 20-21, pp.81-92, 2015.

[23] C. Aggarwal, and N. Li, "On node classification in dynamic content-based networks," Proc. of the 2011 SIAM International Conference on Data Mining (SDM'11), Phoenix, AZ, USA, Apr. 28-30, pp.355-366, 2011.

[24] X. Cui and A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," Genome Biology, Vol. 4, No. 210, pp.1-10, 2008..

[25] S. Bhagat, G. Cormode, and S. Muthukrishnan, "Node classification in social networks," in: Aggrawal, C. (Eds), Social Network Data Analytics. Springer Science+Business Media, LLC., US, pp.115-148, 2011.

[26] P. Melville, "Creating diverse ensemble classifiers. Technical Report, University of Texas, " 2003.

[27] T.G., Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," Machine Learning, 40, pp.139–157, 2000..

[28] G. Biau, L. Devroye, G. Lugosi, "Consistency of random forests and other averaging classifiers," Journal of Machine Learning Research, 9, pp. 2015-2033, 2008.

[29] J. Yun, X. Guocheng,, C. Na, C. Shan, "A New gene expression profiles classifying approach based on neighborhood rough set and probabilistic neural networks Ensemble," In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013, Part II. LNCS, 8227, pp. 484-489, 2013.