Comparing Multifunctionality and Association Information when Classifying Oncogenes and Tumor Suppressor Genes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Gene prediction is an important aspect of understanding cancer pathways. If, for example, an unknown gene is classified to have oncogene or tumor suppressor gene functions, then it would be a candidate for further investigation. In this paper we attempt to classify whether genes are oncogenes or tumor suppressor genes, in other words predicting if proteins have cancer related functions. We show that using simple multifunctionality properties of a gene is a better classifier than complex machine learning such as support vector machine (SVM) and random forest (RF). Since oncogenes and tumor suppressor genes are data sets with very few samples compared to non-cancer related genes, there is a class imbalance, which is solved by performing over-sampling to produce better classification.

1 Introduction

The human genome contains more than 3 billion base pairs, but there are only 20,000-25,000 genes [1]. Genes are DNA segments that carry genetic information, which specify all the proteins that are produced within a gene; genes can produce one or multiple proteins [1]. Proteins perform all the duties that are necessary for a cell to survive [1].

Mutations, changes within a gene, may cause a protein to be unable to perform its originally assigned function(s) [1]. Mutations in oncogenes and tumor suppressor genes can have a dramatic impact on determining whether a cell is transformed from a normal cell to a cancerous cell [2]. Oncogenes have the potential to cause cancer, whereas, tumor suppressor genes are considered the guardian of our cells since they protect the cell from one possible path to cause cancer [2].

A protein may bind directly and perform its assigned function, or multiple proteins may interact either directly or indirectly in order to perform a task [2]. Akt is an oncogene and the Akt signalling pathway has many proteins that interact in order to perform the following functions: allow a cancer cell to uncontrollably divide to produce continuous amount of cancer cells, stop cells from killing themselves even if they are cancerous cells, and increase the NF-kB oncogene [2].

Producing a drug to stop cancerous pathways is very complex since it requires knowing the functions of the proteins that the drug is going to bind to, along with how it will affect the rest of the pathway and the interactions between proteins [3]. If we just produce a protein that will bind to Akt, this will inhibit functions that are actually necessary for normal cell life, which is neither acceptable nor helpful [3].

The ability to know a gene's function within a pathway is vital for understanding how the pathway works and, in the long run, being able to produce an efficient drug [3]. The problem is that it's impossible and impractical to perform every experiment to determine the function(s) of each gene so the ability of predicting a gene function is essential [1]. A popular way to determine the function of a gene is by using associations between their proteins through a protein-protein interaction network (PPIN), which lists all of the interactions between proteins [4]. The underlying idea is that proteins that interact tend to have the same function [4]. Recently, a simpler method with better performance to determine a gene's function was proposed, which analyzes the multifunctionality, the number of functions that a gene has, instead of associations [4].

In this paper, we will determine if multifunctionality, rather than association information, of a gene is a better classifier between a cancer related gene and a non-cancer related gene.

2 Data Sets

There were four data sets used in this paper. The first two data sets are lists of the known oncogenes and tumor suppressor genes [5]. There are 74 known oncogenes and 16 known tumor suppressor genes, but for this paper only 46 oncogenes and 14 tumor suppressor genes that were present in the PPIN were used [5]. The third data set contains node degrees for a list of genes, which was used to determine the multifunctionality of a gene [4]. The fourth data set was the PPIN for 15439 genes, which was where the association information for the features were extracted and were used in training and validating the SVM and random forest [4].

3 Methods

3.1 Features

A PPIN is interpreted as a graph where the nodes are the genes and the edges are the interactions [6]. The following are four prominent features derived from a PPIN that classify between cancer related genes (oncogenes and tumor suppressor genes) and non-cancer related genes [6]:

1) Degree of a node, the number of genes that a gene directly interacts with

2) Number of cancer related genes within one interaction (a node's immediate neighbors)

3) Number of cancer related genes within two interactions (a node's immediate neighbors and the immediate neighbor's neighbors)

4) Average distance (number of interactions) to all cancer related genes

An example for feature extraction can be seen in Figure 1. In this paper, only features 1-3 are used since using a breadth first search within a graph of 15439 genes to extract feature 4 is too computationally intense.

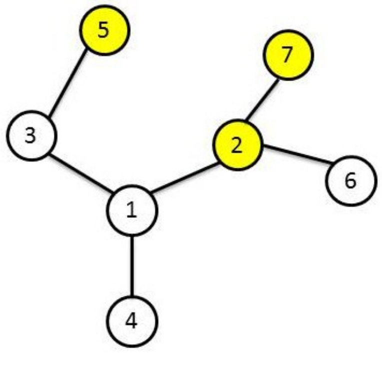

Figure 1: Example for feature extraction from the PPIN. When considering gene 1, the node degree (feature 1) is 3, the number of cancer related genes within one interaction (feature 2) is 1, the number of cancer related genes within 2 interactions (feature 3) is 3 and the average distance to all cancer related genes (feature 4) is (1+2+2)/3=1.6667.

3.2 Classification

There are five steps to classify between cancer related genes and non-cancer related genes using multifunctionality [4]. The first step is to make a list of how the genes rank in terms of multifunctionality; an existing list was used for this paper [4]. Secondly, all the genes are labelled as cancer related genes that are between the highest ranked gene and the lowest ranked cancer related gene [4]. The third step includes calculating the true positive (TP) rates and false positive (FP) rates [4]. The fourth step is to repeat the second and third step to obtain a variety of TP rates by changing the labelling system (i.e. label all genes from the highest ranked gene to the ith lowest cancer related gene) [4]. The final step is to determine the area under the ROC curve, TP rates vs. FP rates, which is shown to be equivalent to correct classification rates (CCR) [4].

Association information was used to train and test a linear support vector machine (SVM) from the Bioinformatics toolkit in Matlab and a random forest (RF) with 500 trees [7]. For both SVM and RF, using the full set of non-cancer related genes caused an out of memory error. Random under-sampling, which randomly chooses a subset of data, was used to solve this problem [8]. We chose to do three separate experiments, involving randomly under-sampled subsets from 15439 to 1000, 500 and 100 non-cancer related genes. The validation was performed by taking the average of thirty 2-fold cross validations for each SVM and RF using the combination of all three features.

Additionally, to compensate for such a small oncogene and tumor suppressor gene sample, we used an over-sampling technique [8]. There are simple over-sampling algorithms that just repeat the low quantity samples until too high of a cost value is reached, but this produces extreme overlapping [8]. A preferred approach, used in this paper, is to apply Synthetic Minority Over-sampling Technique (Smote). Smote produces more examples for the underrepresented group by interpolating between multiple examples from the group that have similar features [8]. This avoids the overlapping issue, which allows the boundaries for the minority group to be larger and more certainty in classification [8].

4 Results

Choosing features can determine if the classification will be successful or a failure. The average value to expect for each feature in each data set is shown in Table 1. Non-cancer related genes have a low mean value for all three features, whereas oncogenes have the highest mean value for all features. The feature extraction for the full PPIN was the most computationally expensive with a runtime ranging from 15-30 minutes for each.

In Figure 2, the ROC curve for classifying oncogenes and tumor suppressor genes using multifunctionality are very similar, which means they have the similar classification. The area under curve (AUC) for ROC curve was 0.4930 for classifying oncogenes and 0.4962 for classifying tumor suppressor genes. These values are poor classification rates and indicate that the oncogenes and tumor suppressor genes are located within the top half of the ranked node degree list.

Table 1: Mean of each feature for the three different data sets. Oncogenes have the highest mean value for all the features and non-cancer related genes tend to have lower mean values for the features. Random under-sampling will be performed on the non-cancer related gene set, while over-sampling will be performed on the oncogenes and tumor suppressor genes. In both cases, the mean values for the features will give an estimate of what to expect.

	Number of Genes	Mean of Feature 1	Mean of Feature 2	Mean of Feature 3
Oncogene	46	32.0870	0.0652	3.1739
Tumor suppressor gene	14	58.0714	0	0.2143
Non-cancer related gene	15439	3.8019	0.0069	0.7171

The True Positive (TP) rate for oncogenes represents the number of properly classified oncogenes. Accuracy is defined as TP + True Negative (TN) rate, which is the properly classified oncogenes and non-cancer related genes. As seen in Table 2, the accuracy for the classification of oncogenes is extremely high, whereas the TP rate is extremely low. Also, as we decreased the number of non-

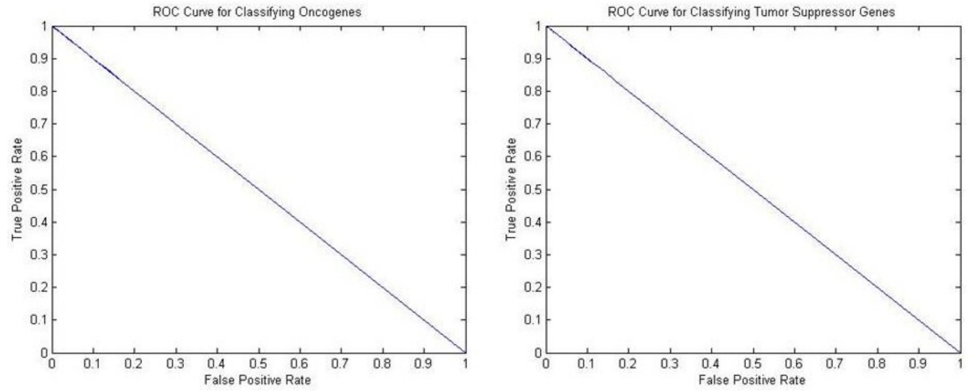

Figure 2: ROC curve for the classification of oncogenes and tumor suppressor genes using multi-functionality. The AUC of the ROC curve for both are approximately 0.5, which means that both oncogenes and tumor suppressor genes are within the top half ranked for node degree.

Table 2: Classification of oncogenes using SVM and RF. Both accuracy and true positive (TP) rates are shown because the class imbalance causes the accuracy measurement to be misleading. There is a high accuracy but very few oncogenes are actually being properly classified. There are three different size data sets that non-cancer related genes were under-sampled to. Accuracy is known as the TP rate + true negative rate. The accuracy and TP values were taken as an average of thirty 2-fold validation.

	Accuracy			True Positive Rate		
Number of non-cancer related genes	1000	500	100	1000	500	100
SVM	0.9584	0.9211	0.7822	0.1304	0.2493	0.4159
Random forest	0.9565	0.9221	0.7904	0.0543	0.2304	0.4884

Table 3: Classification of tumor suppressor genes using SVM and RF. The class imbalance causes misleading accuracy values since there is a high accuracy, but low true positive (TP) rates shown. Accuracy is known as the TP rate + true negative rate. The accuracy and TP values were taken as an average of thirty 2-fold validation. There are three different size data sets that non-cancer related genes were under-sampled to.

	Accuracy			True Positive Rate		
Number of non-cancer related genes	1000	500	100	1000	500	100
SVM	0.9847	0.9779	0.9078	0.1250	0.4063	0.5188
Random forest	0.9825	0.9694	0.8819	0	0.0187	0.4437

cancer related genes, the accuracy decreases and the TP rate increases. The exact same pattern is seen for tumor suppressor genes in Table 3.

After over-sampling to acquire more oncogenes and tumor suppressor genes, the RF has the better average true positive rate value, as seen in Tables 4 and 5. The RF even had a high of 0.75 which is above the value that is considered a good classification. The number of oncogenes or tumor suppressor genes after over-sampling depends on the size of non-cancer related genes and the cost matrix for Smote [8].

5 Discussion

In regards to the non-cancer related genes feature, the mean of all the features are low, which is to be expected since the set is so large the variance is going to be huge. Since we are randomly selecting

Table 4: Classification of oncogenes performed on SVM and RF on over-sampling. The increase of oncogenes from over-sampling depends on the size of the non-cancer related genes set. Each column explains how many non-cancer related genes were used, the number of oncogenes were used after performing over-sampling and the classification values for SVM and random forest.

	Accuracy		True Positive Rate	
Number of non-cancer related genes	300	100	300	100
Number of oncogenes after over-sampling	120	43	120	43
SVM	0.8669	0.7936	0.4667	0.4826
Random forest	0.8220	0.8009	0.6609	0.6667

Table 5: Classification of tumor suppressor genes performed on SVM and RF on over-sampling. The increase of tumor suppressor genes from over-sampling depends on the size of the non-cancer related genes set. Each column explains how many non-cancer related genes were used, the number of oncogenes were used after performing over-sampling and the classification values for SVM and random forest.

	Accuracy		True Positive Rate	
Number of non-cancer related genes	300	100	300	100
Number of tumor suppressor genes after over-sampling	113	43	113	43
SVM	0.9184	0.8862	0.6250	0.6500
Random forest	0.8829	0.8560	0.6563	0.7500

non-cancer related genes, this gives a good estimate on what to expect. Feature 1 seems to be the best, since it contains the largest distance between the groups.

The number of known non-cancer related genes heavily outnumbers the known oncogenes and tumor suppressor genes; this is known as a class imbalance. We further reduced the already small data set by only using 46 of the total 74 oncogenes and 14 of the total 16 tumor suppressor genes because they were not present in the PPIN.

This class imbalance between 14 or 46 cancer related genes and 15439 non cancer related genes causes misleading conclusions when comparing accuracy. During SVM and RF the method labelled all genes as non-cancer related genes, which resulted in a very high accuracy since they are a significantly larger group. As a result, a very low positive prediction for cancer related genes was produced.

As seen in Tables 2 and 3, multifunctionality is a better classifier since no potentially important data was discarded in random under-sampling and a higher number of true positive classification rates was produced. The AUC for ROC curves is approximately 0.50 for both oncogenes and tumor suppressor genes, which is lower than what is considered good at 0.70.

Genes that have many gene functions should be considered with higher priority when classifying a gene function since it is a much higher probability that a gene that has many functions will have another function [4]. Ranking genes by their multifunctionality to predict a genes function is a powerful method, since it is easy, quick and has a higher classification rate.

The TP Rate for SVM and RF using over-sampling increased by 15-40 percent; this can be seen when comparing Tables 2-5. We were unable to perform over-sampling on the multifunctionality data since it was obtained from a published paper.

6 Conclusion

In this work, we analyzed different classification methods that determine if a gene is cancer related. Gene multifunctionality outperforms association information used in SVM and random forest for classification on class imbalance data. The best classification results were obtained when performing SVM with over-sampled data using Smote to resolve the class imbalance. The process of resampling for more cancer related genes, in order to correct the class imbalance, was not able to be performed on the node degree list for multifunctionality since the list was obtained from another paper and cannot be manipulated. We cannot determine which method would perform the best on the over-sampled data but it is clear that using multifunctionality outperforms SVM or RF and using Smote on class imbalance data does improve the classification.

Future work would include performing experiments to determine the improvement of classification on multifunctionality using over-sampled data. Additionally, we would like to investigate further into different over-sampling techniques, features selection and data sets.

References

[1]Sullivan, L. & Johnson, R. & Mercado, C. & Terry, K. (2009) *Human Genome: The SAGE Glossary of the Social and Behavioral Sciences.* SAGE Publications, Inc.

[2] Carnero, A. (2010) The PKB/AKT pathway in cancer *Current Pharmaceutical Design* **16**(1):34-44

[3] Xie, L. & Evangelidis, T. & Xie, L. & Bourne, P. (2011) Drug Discovery Using Chemical Systems Biology: Weak Inhibition of Multiple Kinases May Contribute to the Anti-Cancer Effect of Nelfinavir *Public Library of Science***4**(7):e1002037

[4] Gillis, J. & Pavlidis, P. (2011)The Impact of Multifunctional Genes on ”Guilt by Association” Analysis. *PLoS ONE* **6**(2):e17258.

[5] http://www.cancerquest.org/oncogene-table.html

[6] Li, Y. & Patra, C. (2008)Selection of features from protein-protein interaction network for identifying cancer genes.*Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference onl*, pp.1707 -1712.

[7] http://code.google.com/p/randomforest-matlab

[8] Batista, G. & Prati, R. & Monard, M. (2004) A study of the behavior of several methods for balancing machine learning training data.*SIGKDD Explor. Newsl* **6**(1):20-29.

[9] Gillis, J. & Pavlidis, P. (2011)The role of indirect connections in gene networks in predicting function. *Bioinformatics*(10):1093.