Searching for Patterns in Imbalanced Data: Methods and Alternatives with Case Studies in Life Sciences

A. Fazel Famili

Information and Communications Technologies, National Research Council Canada,
1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
fazel.famili@nrc-cnrc.gc.ca

Abstract. The prime motivation for pattern discovery and machine learning research has been the collection and warehousing of large amounts of data, in many domains such as life sciences and industrial processes. Examples of unique problems arisen are situations where the data is imbalanced. The class imbalance problem corresponds to situations where majority of cases belong to one class and a small minority belongs to the other, which in many cases is equally or even more important. To deal with this problem a number of approaches have been studied in the past. In this talk we provide an overview of some existing methods and present novel applications that are based on identifying the inherent characteristics of one class vs the other. We present the results of a number of studies focusing on real data from life science applications.

Keywords: Knowledge discovery, imbalanced data, gene expression data.

1 Introduction

Over the past 10-20 years, we have noticed an extensive emergence of tools and technologies to design experiments, automatically monitor, collect and warehouse large amounts of data, in many domains such as life sciences and industrial processes. This has become the prime motivation for pattern discovery and machine learning research and development. Examples of unique problems in this domain are situations where the data is imbalanced. The class imbalance problem corresponds to situations where majority of cases (samples) belong to one class and a small minority belongs to the other, which in many cases is equally or even more important. The goal is still to learn from this data and discover patterns that are useful and unknown to the producers of this data or domain experts. Most algorithms are overwhelmed by the majority class and ignore the minority class since the traditional classifiers focus more on minimizing the overall error rate instead of paying special attention to the minority class. This could result in classifying all the data into the majority class in order to achieve higher accuracy. As an example, decision trees tend to over-generalize the class that is represented by most of the examples in the data. This obviously creates a major problem.

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 159–166, 2014.

Analyzing imbalanced data sets is a great challenge in machine learning, data mining and knowledge discovery in real world applications. Examples are, spotting unreliable telecommunication customers, detection of oil spills in satellite radar images, learning word pronunciations, text classification, risk management, information retrieval and filtering tasks, network monitoring and intrusion detection, fraud detection, shuttle system failure, earthquakes, nuclear explosions and helicopter gear-box fault monitoring as well as medical diagnosis (e.g. rare disease and rare genes mutations) [10]. Although we have not seen explicitly stated anywhere, ratios of 2:1, 5:1, and 10:1 have often been used in experiments under the category of imbalanced data sets. So ratios of 2:1 or better will not be regarded as imbalanced by many.

Many traditional algorithms in machine learning and data mining problems assume that the target classes share similar prior probabilities. Therefore, these algorithms are sometimes overwhelmed by the majority class and ignore the minority class since the traditional classifiers focus more on minimizing the overall error rate instead of paying special attention to the minority class. This could result in classifying all the data into the majority class in order to achieve higher accuracy. For instance, decision trees tend to over-generalize the class with the most examples, while a Naïve Bayes method requires enough data for the estimation of the class-conditional probabilities. Case-based methods, in this respect, have been mentioned in the literature as the most appropriate choice since they work on a prototype representation of classes and classify new samples according to the ones seen so far [5]. In particular, in these works it is argued that the poor performance of the classifiers produced by the standard machine learning algorithms on imbalanced datasets is mainly due to the following three factors: accuracy, class distribution and error cost. This is due to the fact that they are rarely well satisfied in real world applications. Most algorithms behave badly when the datasets are highly imbalanced [10].

Since the emergence of microarray technology in 1995, an enormous number of gene expression pattern recognition methods have been introduced to improve the accuracy of medical diagnosis, prognosis among which are several success stories, including some commercial applications and diagnostic tools [9], [24], [25]. It is very crucial to include in the studies rare but important cases (normally not too many available) in microarray data analysis for the disease understanding and patients' subsequent treatment (accurate diagnosis, better personalized medical care). However, either because of the high costs associated to obtain the data or the disease/case being rare (such as sub type of certain diseases, treatment related acute myeloid leukemia, brain cancer, ovary cancer, pancreas cancer), some of today's high throughput (e.g., gene expression) data sets could be highly imbalanced. As a result, most traditional classifiers (i.e. data mining techniques) are not able to effectively deal with the problem of imbalanced datasets and this has become the new challenge in microarray (omics) data analysis, when developing computerized diagnostic applications.

2 Related Work

There are at least 2 groups of papers that seem to be relevant to this research; (i) papers that discuss the analysis of imbalanced data and their proposed approaches and

(ii) papers on analyzing small/imbalanced data sets in high throughput genomics. There have been a number of attempts all investigating the class imbalance problem [4], [12], [16]. To tackle the problem, different methods have been proposed and experimented.

Changing class distributions by under-sampling the major-represented classes [17], [18], [20] or over-sampling the under-represented class [4], [11], [23] to make the classes balanced is one approach. In addition, there are several heuristic over-sampling methods mainly based on SMOTE [4], [11], which generates synthetic examples of the under-represented class in order to over-sample such a class. However it is reported that the under sampling could discard potential useful data that could be important for the classifier [10]. It is also known that over-sampling can increase the possibility of over fitting since it makes exact copies of the minority class examples. Treating the imbalance problem at the algorithmic level is another way. For example, Provost and Fawcett [21] built a hybrid classifier by adjusting the costs of the various classes, the probabilistic estimate at the tree leaf (when building decision trees), and the decision threshold. Other researchers have also combined the above two methods to deal with imbalanced problem [13], [27].

Both sampling techniques and algorithmic methods may not work well for class imbalance problems that deal with high dimensional data. This is the exact case in analyzing high throughput genomics data which is what we are discussing in this paper. Zheng *et al* [28] proposed a feature selection framework, which selects features for positive and negative classes separately and then explicitly combines them. This approach shows simple ways of converting existing measures so that they separately consider features for negative and positive classes. Van Der Putten and Van Someren [26] used bias-variance decomposition to analyze the COIL 2000 dataset and reported that to avoid over-fitting choosing proper feature selection is even more important than the choice of the learning method.

Another area of related research is the ensemble learning where these methods have been extensively used to handle class imbalance problems. These methods combine the results of many classifiers. Their successes can be attributed to the fact that their base learners usually are of diversity in principle or induced with various class distributions [10]. A number of past researches have reported that ensembles of base learners exhibit substantial performance improvement over single base learners [6]. The resulting classifiers, referred to as ensemble classifiers, are the aggregation of classifiers whose individual decisions are combined by weighted or unweighted voting to classify new samples. Bagging introduced by Breiman [1] and boosting proposed by Freund and Schapire [7] are well-known and popular representatives of this methodology. Random forest is also a popular ensemble technique developed by Breiman [2], which has shown successful results in dealing with imbalanced data sets.

Although our main objective in this research has not been to develop a new method to deal with imbalanced data sets, we have explored an ensemble approach to deal with this problem so that we can properly analyze high dimensional microarray-based data sets targeting disease classification.

3 Data Sets Used

We used two gene expression datasets from public domain, which were published by Bullinger *et al.* [3], and Klein *et al.* [15]. They are available at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/, with accession number GSE425 and GSE15434, respectively).

The first data set by Bullinger contains the gene expression of 116 de novo AML (Acute Myeloid Leukemia) and 10 secondary AML patients, which were obtained using cDNA microarrays platform. Klein's data, our second data set, contained the gene expression of 251 AML patients with normal karyotype, obtained using Affymetrix Human Genome U133 Plus 2.0 Array platform. In addition, we used a complementary data set which was a bit bigger where we had access to 28 t-AML samples, plus 367 AMLs. The data size was therefore 435 samples in total with close to 40000 attributes (probes) [14].

4 Our Approach

This section provides an overview of the proposed approach (Fig. 1). Microarray data containing a large class (L) and a small class (S) are first passed through a basic data preprocessing step which includes data normalization, data filtering, and missing data imputation based on domain expertise and additional research. The next step in our approach is the dimension deduction/feature selection. Here we consider only genes that are differentially expressed between the two classes using a two sample *t*-test. We randomly select *r* samples from the large class (L). The possible number of combinations/selections is $\boldsymbol{C_L^r = \frac{L!}{(L-r)!r!}}$ where L is the total number of samples. These *r* randomly selected samples from the large class combined with the small class are used in a knowledge discovery method called "Discover & Mask" [19] with specific accuracy and a predefined threshold for maximum number of models generated by the program. This procedure should continue as many (N) times as possible according to the size of the data. In each run a number of tree models will be generated, where each tree model consists of one or more genes (nodes) with corresponding thresholds to discriminate between the classes. These genes are considered informative as to the classification task. This process is repeated as many times as possible and a list of genes are collected from each of these runs. The genes with the higher number of occurrences in the run give us more confidence: no matter how the data are selected. We can claim that the identified genes would be a better selection in terms of discriminating classes. Finally the high ranking genes are validated using public databases and domain knowledge.

5 Results

We evaluated the performance of the proposed methodology through its application to the datasets described in Section 4. Table 1 shows the experiment parameters and

results. Bullinger's "well-measured and highly variable" subset of 6283 probes was adopted as the starting point. For Klein's data, we used the whole data (54675 probes) and a subset of 19070 genes, which is the pre-processed subset with MAS5.0 using brainarray CDF file. We used a two sample t-test (p value cutoff <=0.05 and fold change 0.58) to identify lists of differentially expressed probes (or genes) from these datasets. 290, 1611 and 3735 probes (or genes) were identified respectively from the three datasets. To apply the Discover & Mask algorithm, we set N to 1000, 2000 and 5000, respectively for the three datasets. Further we set the accuracy to 95% and selected a maximum of models to 10 (for each run). Top 100 probes (or genes) from each result sets were selected, which resulted in 88, 100 and 86 known/annotated probes (or genes) for further investigation.

It is interesting to notice that there are only few common genes identified from different datasets (see Table 1). The reason could be i) the difference in the different microarray platforms, ii) the effect of data pre-processing process. Consolidation of these results could trim down the effect of the differences and produce complementary outcomes. This shows that one cannot reply on a single approach to analyse microarray data.

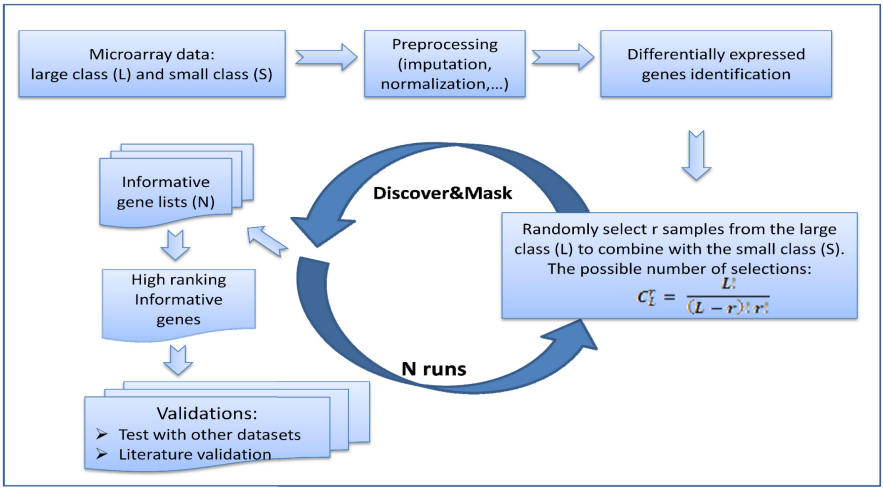

Fig. 1. The proposed approach for handling imbalanced microarray data

We validated part of the results of our approach using public databases such as PubMed, GeneCards [22], which is curated according to the information from biomedical literature databases (OMIM, SWISS-PROT, Genatlas, GeneTests, GAD, GDPInfo, Bioalma, Leiden, Atlas, BCGD, TGDB, and/or HGMD). We have some interesting findings shown in Table 1: (i) some of the genes are associated with cancer treatment drugs. API5 (apoptosis inhibitor 5) is one of the examples.

Table 1. Experiment parameters and results

	Bullinger's Data	Klein's Data	
Data Used	subset	subset	whole data
- Number of probes (p)/genes (g) start with	6283 (p)	19070 (g)	54675 (p)
t-test results	290	1611	3735
Discover & Mask:			
- running times (N)	1000	2000	5000
- accuracy threshold	95%	95%	95%
- maxi # of j48 Models	10	10	10
Random sample size (r)	30	50	50
Number of probes/genes with annotation in top 100 significant probes(genes)	88	100	86
Common genes between datasets	2	2	
Common genes between datasets		6	6
Common genes between datasets	0		0
Interesting Genes	TCF4, API5, FOXO3, HOXB7	SEPTIN9	FCAR, KDM3B

6 Conclusions

Therapy-related myeloid leukemia (t-AML) is a well-recognized clinical syndrome occurring as a late complication following cytotoxic therapy. With the existence of microarray technology, it is feasible to identify gene expression profiles to be used for the stratification of t-AML from denovo AML or normal patients. However:

- A much larger sample size would be needed for the analysis to be valid.
- For identifying t-AML predisposition development conditions, patient samples must be acquired prior to treatment.

This study focuses on analyzing imbalanced gene expression data from a combination of t-AML and AML data sets. Considering several possibilities to handle imbalanced data sets we discuss different approaches that have produced some interesting results for identifying informative genes, while a few have been validated through literature validation to be associated with the disease. We also discuss using data characteristics in order to identify subsets from the majority class in imbalanced situations. As a future work we plan to investigate other methods to analyze the same data.

References

1. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R.F., Tibshirani, R., Döhner, H., Pollack, J.R.: Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. N. Engl. J. Med. 350, 1605–1616 (2004)
4. Chawla, N.V., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Colantonio, S., Little, S., Salvetti, O., Perner, P.: Prototype–Based Classification in Unbalanced Biomedical Problems. In: Montani, S., Jain, L.C. (eds.) Successful Case-based Reasoning Appl. SCI, vol. 305, pp. 143–163. Springer, Heidelberg (2010)
6. Dahinden, C.: An improved Random Forests approach with application to the performance prediction challenge datasets. In: Guyon, I., et al. (eds.) Hands on Pattern Recognition. Microtome (2009)
7. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
8. Fuller, J.F., McAdara, J., Yaron, Y., Sakaguchi, M., Fraser, J.K., Gasson, J.C.: Characterization of HOX gene expression during myelopoiesis: role of HOX A5 in lineage commitment and maturation. Blood 93(10), 3391–3400 (1999)
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439), 531–537 (2009)
10. Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G.: On the Class Imbalance Problem. In: Proc. of 4th International Conference on Natural Computation, Jinan, October 18-20, pp. 192–201. IEEE, Los Alamitos (2008)
11. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005)
12. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Journal Intelligent Data Analysis Archive 6(5) (2002)
13. Joshi, M.V., Kumar, V., Agarwal, R.C.: Evaluating boosting algorithms to classify rare cases: comparison and improvements. In: First IEEE International Conference on Data Mining, pp. 257–264 (2001)
14. Kharas, M.G., Lengner, C.J., Al-Shahrour, F., Bullinger, L., Ball, B., Zaidi, S.: Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. Nature Medicine 16(8), 903–908 (2010)
15. Klein, H.U., Ruckert, C., Kohlmann, A., Bullinger, L., Thiede, C., Haferlach, T., Dugas, M.: Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. BMC Bioinformatics 10, 422 (2009), doi:10.1186/1471-2105-10-422
16. Kubat, M., Holte, R., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning 30, 195–215 (1998)
17. Kubat, M., Matwin, S.: Addressing the curse of imbalanced data set: One sided sampling. In: Proc. of the Fourteenth International Conference on Machine Learning, pp. 179–186 (1997)

18. Liu, X.Y., Wu, J., Zhou, Z.: Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(2), 539–550 (2009)
19. Ouyang, J., Famili, F., Xu, W.: An Approach to Automated Knowledge Discovery in Bioinformatics. In: Li, D., Wang, B. (eds.) Proceedings of the Conference on Artificial Intelligence and Innovations (AIAI 2005). IFIP, vol. 187, pp. 593–600. Springer, Boston (2005)
20. Padmaja, T.M., Dhulipalla, N., Krishna, P.R., Bapi, R.S., Laha, A.: An unbalanced data classification model using hybrid sampling technique for fraud detection. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 341–348. Springer, Heidelberg (2007)
21. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning 42, 203–231 (2001)
22. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D.: GeneCards: Integrating information about genes, proteins and diseases. Trends Genet. 13(4), 163 (1997)
23. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proc. 24th Int. Conf. Mach. Learn., Corvallis, OR, pp. 935–942 (2007)
24. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., et al.: A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. 347 (2002)
25. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)
26. Van Der Putten, P., Van Someren, M.: A biasvariance analysis of a real world learning problem: the coil challenge 2000. Machine Learning 57(1-2), 177–195 (2004)
27. Weiss, G.M.: The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning. Ph.D. Dissertation, Department of Computer Science, Rutgers University, New Brunswick, New Jersey (2003)
28. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. SIGKDD Explorations 6(1), 80–89 (2004)