Disease Prediction Based on Prior Knowledge

Gregor Stiglic, Igor Pernek, Peter Kokol
Faculty of Health Sciences
University of Maribor
Zitna ulica 15, 2000 Maribor, Slovenia
{gregor.stiglic, igor.pernek, kokol}@uni-mb.si

Zoran Obradovic
Center for Information Science and Technology, Temple University
Philadelphia, PA 19122, USA
zoran.obradovic@temple.edu

ABSTRACT
Increasing demand for digitalization of Electronic Health Records results in increased demand for effective data mining solutions. In this study we enhance the classical Support Vector Machine - Recursive Feature Elimination (SVM-RFE) approach to optimally estimate disease risk from hospital discharge record data. Our approach is based on incorporating prior knowledge from human disease networks extracted from hospital discharge historical data and lowering the burden of building classifiers from huge amounts of data. To predict future risk of hospitalization based on highly imbalanced and 11,170 dimensional hospital discharge data consisting of nearly 7 million records collected in year 2008, we adopt a knowledge representation from complex systems and a feature selection technique used in bioinformatics. Our out of sample results on year 2009 dataset of similar size provide evidence that the proposed method is beneficial in cases where the classical SVM-RFE model is unstable. When using the new method we demonstrate that stability is improved in cases where one aims to remove large batches of features in a single iteration.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data mining

General Terms
Algorithms

Keywords
Feature Selection, Classification, Risk Estimation, Disease Networks, Support Vector Machines.

1. INTRODUCTION
Much work has been done in the past years by healthcare organizations to bring health data into digital form. With the increased acceptance of electronic health records, we can observe the increased interest in the application of data mining approaches in this field. Large datasets containing patient information are becoming available and physicians are able to compare their current patients to patients with similar diagnoses to decide on appropriate treatments. In recent years we have witnessed an increased amount of studies focusing on complexity of relations and co-occurrence of multiple diseases.

A recent study by Steinhaeuser and Chawla [16] points out that our health care system is mostly reactive, meaning that we have become proficient at diagnosing diseases and developing treatments to cure them or prolong the life of a patient. However, we should now put more focus on proactive care, aiming especially at prediction of disease-related risks before they actually happen, and guiding the patient to avoid them, instead of just curing them. Large amounts of available clinical data can be used to achieve this goal.

One of the first approaches using a large number of patients to construct disease related networks was a study by Hidalgo et al. [7] where authors demonstrated the usefulness of human disease networks in studying the properties of co-occurring diseases. The study also demonstrates the potential of phenotypic data in the form of a human disease network to complement genotypic and proteomic datasets that were extensively studied and analyzed in the past. The main contribution of the mentioned study that is also used in our research is the introduction of the metric for quantification of comorbidity relationships.

The integration of networks in improving diagnostic and prognostic methods has recently been very popular in bioinformatics, where gene and protein networks have been used in feature selection problems. Integration of biological prior knowledge (e.g. protein-protein interaction (PPI) networks) has been used to improve the performance of the predictive algorithms and to increase the stability of biomarker signatures [4]. Taking into account the high dimensionality of data in medical record datasets, especially in patient records containing more than ten thousand diagnosis and procedure codes, the problems in bioinformatics and healthcare informatics seem to have a lot of similarities.

An approach that exploits the information obtained by constructing an interconnected network of human diseases in order to improve the performance of a classification algorithm for disease risk based on comorbidity information is presented in this paper. The following section presents the basics of using prior knowledge for classification tasks in the bioinformatics domain. The idea from bioinformatics is applied to hospital discharge data and the disease risk estimation problem with our experimental setup presented in Section 3. Section 4 describes the results of the experiments and is followed by the last section with conclusions and future work.

HI-KDD'12, Aug 12, 2012, Beijing, China.

2. BACKGROUND

Predictive modeling based on hospital discharge data has only recently attracted the attention of medical informatics and knowledge discovery communities. Moturu et al. [11] use hospital discharge data for early identification of high cost patients. Their study compared five different classification models to classify patient records where the total cost of a hospital stay exceeded a predefined threshold that defined a record as "high-cost". They tested a variety of popular classification algorithms and selected the five most appropriate classifiers for the task: AdaBoost (with 250 iterations of a Decision Stump classifier), Logit-Boost (also with 250 iterations of a Decision Stump classifier), Logistic Regression, Logistic Model Trees, and the Support Vector Machine (SVM).

Another study that used International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) codes to predict disease risks was presented by Davis et al. [5]. Their approach is based on collaborative filtering and uses temporal dimension of discharge data to predict future risks. The data used by Davis et al. comprises of more than 13 million patient records, but is limited to populations older than 65 years. A similar risk prediction study was conducted by Khalilia et al. [9] where the NIS datasets from 2005 were used to classify records based on grouped diagnosis codes. This study compared four different classifiers and found out that the Random Forest classifier produced the best results in disease classification.

In our study, we focus on the feature selection problem for high-dimensional hospital discharge classification problems. The next section reviews the related work that was done on high-dimensional datasets in bioinformatics.

2.1 Related Work

Various methods from machine learning have been used for identification of gene expression signatures in the past [20, 21]. However, different studies demonstrate [15, 17] that current feature selection methods usually lack stability and biological interpretability in high-dimensional data. To solve this problem, one may note a growing interest in methods that try to integrate the prior knowledge from biological networks to improve the stability, interpretability, and, consequently, classification performance of the previously proposed feature selection methods.

In general, there are two groups of methods integrating network knowledge – i.e. network centric and data centric [4]. The first group focuses on mapping gene expression data onto a network and uses techniques from network analysis to select the important genes. On the other hand, the second group focuses on machine learning techniques where prior knowledge from biological networks is used to bias the feature selection process toward strongly connected genes. The approach used in this study, called SVM-RRFE (Reweighted Recursive Feature Elimination), was proposed by Johannes et al. [8] and belongs to the group of data centric methods.

A well-known GeneRank algorithm [10] was used by Johannes et al. to couple SVM-RFE feature selection method with a network of connected features (e.g. PPI network). Inspired by Google's PageRank [12], GeneRank calculates importance of genes in supervised learning problems as a combination of their fold-change (simple univariate feature selection technique) and their centrality measure in the network. They use the GeneRank measure of gene importance in each iteration of SVM-RFE to re-adjust the SVM decision hyperplane. In this way, authors try to integrate three different feature importance measures (SVM, fold-change and network) into a single feature selection method (SVM-RRFE), thus significantly improving the stability and interpretability of selected features.

2.2 Classification and Class-imbalanced Data

The problem of class-imbalance occurs when a classification algorithm is confronted with a dataset containing a small number of positive (i.e. predicted diagnosis code) samples and a much larger number of other samples. For example, E.Coli bacterial infection is present in less than 0.5% of hospitalization records. Most classification algorithms tend to ignore the samples of the rare class and focus on correctly predicting the majority class. This problem is prevalent in most fields of applied machine learning and is also one of the important issues in medical data mining [3]. The problem of imbalanced data can be addressed at the data or algorithmic level. In the first case, one can use different variants of undersampling the majority or oversampling the minority class. On the algorithmic level, the class imbalance problem is addressed by using different variants of cost-sensitive learning or learning from one class rather than two [2]. Most of the studies that use hospital discharge data for classification are using a variant of undersampling the majority class. Usually the ratio of the small to other classes in hospital discharge data varies significantly. Khalilia et al. [9] report the imbalance rates ranging from almost 30% to under 0.1% in disease risk classification on NIS dataset.

Additionally to the problem of building the classifier in unbalanced settings, it is also important to properly measure the classification performance in such settings. The classical classification accuracy should be replaced with a measure that will put more focus on the classification performance for rare positive samples. Tang et al. [18] identified four metrics that are most suitable for evaluation of classifiers on unbalanced data: geometric mean of sensitivity and specificity, area under ROC curve, F-measure and area under precision-recall curve. Similar to other experiments in this field, our study uses the area under ROC curve (AUC) metric.

2.3 Human Disease Networks

There are two basic concepts that are usually used when constructing human disease networks: morbidity, representing the support for a single diagnosis in the given population; and co-morbidity, the support for co-occurrence of two diseases. In our experiments we compare the efficiency of three co-morbidity measures: weight (by Steinhaeuser and Chawla [16]), relative risk, and phi (both by Hidalgo et al. [7]). The weight of the edge, connecting diseases *i* and *j* can be calculated as follows:

$$W_{ij} = \frac{C_{ij}}{M_i + M_j}$$

where C_{ij} is co-morbidity of two compared diseases and M is morbidity or prevalence of a single disease. Weight measure aims to balance the high values for more frequent co-morbidities by dividing their number by a sum of single disease prevalence.

The relative risk measure is similar to weight, but also includes the total number of patients in the population (N) and is defined as:

$$RR_{ij} = \frac{C_{ij}N}{M_i M_j}$$

Relative risk measure is intrinsically biased towards overestimation of relationships between rare diseases and underestimates the co-morbidity of more frequent diseases. This bias can be reduced by introduction of a ϕ-correlation measure, defined as:

$$\phi_{ij} = \frac{C_{ij}N - M_i M_j}{\sqrt{M_i M_j (N - M_i)(N - M_j)}}$$

To focus on relationships between rare diseases that can often be left out by classical feature selection approaches, RR measure was used throughout this study. The implementation and role of the human disease network used in our study is explained in the next section.

3. EXPERIMENTAL SETUP

The proposed classification approach was applied to the Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality [1]. NIS dataset contains hospital discharge records for a stratified sample of approximately 20% of US hospitals. In our study, we used data for the adult population from year 2009 for model evaluation and 2008 for disease network construction, in order to avoid bias in feature selection. Altogether, there are 6,546,273 discharge records in the 2009 dataset and 6,840,196 in 2008.

Each record contains the personal characteristics of the patient, including age, gender, race; administrative information, including length of stay, and discharge status; medical information, including diagnoses, surgical and nonsurgical procedures. Each patient can have up to 15 (2008 dataset) or 25 (2009 dataset) diagnoses. Age group frequencies for both years are compared in Figure 1. The International Classification of Diseases, 9th Revision, Clinical Modification, or ICD-9-CM was used for coding diagnoses. ICD-9-CM coding uses taxonomy of five-digit codes, where the first three digits represent the general diagnosis and are followed by two additional digits describing a more detailed subgroup of the general diagnosis. Using very detailed five-digit codes results in a very complex classification problem due to an extremely high number of samples and features. Altogether, there are 14,315 possible diagnosis codes in our datasets. After removing the codes that were not used in data from year 2009, we could reduce the dimensionality of our datasets to 11,170 features, representing different diagnosis codes that were used at least once.

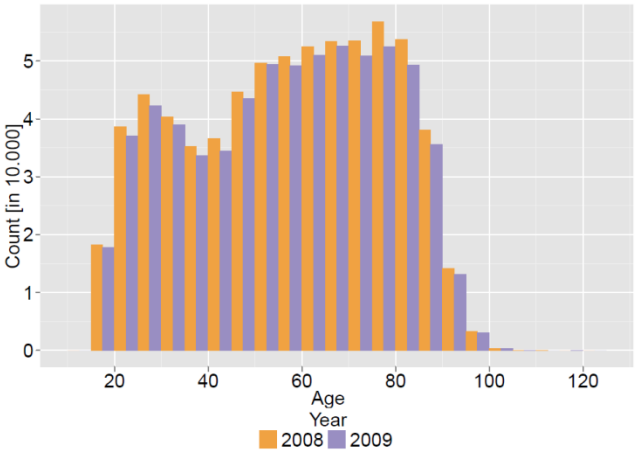

Figure 1. Number of records in datasets by age groups

The NIS dataset used in this study contains 126 clinical and nonclinical features for each hospital discharge record. In addition to demographic information, each record also contains ICD9-CM diagnosis codes and Clinical Classification Software (CSS) codes. CCS is used to collapse ICD9-CM diagnosis codes into more general categories. There are 259 CCS disease categories that could also be used for classification, similar to a study by Khalilia et al. [9]. However, even the authors of the study themselves admitted that classifying diseases using only CCS features may not lead to a valid disease prediction. Therefore, we use the whole feature space of ICD9-CM diagnosis codes for classification in this study. This allows for a broad range of possible classification problems, by selecting any of more than 14,000 diagnosis codes as a target class. More realistically, some very rare diagnosis categories cannot be used due to their extremely low support. Therefore, we focused on prediction of disease categories with prevalence close to 1% and above. We selected 5 diseases in a way that would allow comparison of unbalanced datasets with different ratios of target versus non-target class samples. To obtain the most appropriate target class candidates, we initially ranked all diagnosis codes by their prevalence. The most prevalent disease, with a prevalence of over 36%, is "Essential Hypertension – Unspecified" (diagnosis code 401.9). In our study we compare classification performance for 5 different diagnosis codes obtained by selecting the diagnosis codes closest to a prevalence of 20, 10, 5, 2 and 1%. Table 1 presents some basic information on the selected diagnosis codes.

Table 1. Diagnosis codes used for classification performance estimation with corresponding prevalence for 2009 dataset

Diagnosis code	Description	Prevalence (%)
272.4	Other and unspecified hyperlipidemia	18.42
285.9	Anemia, unspecified	9.29
278.00	Obesity, unspecified	4.95
280.9	Iron deficiency anemia, unspecified	2.00
578.9	Hemorrhage of gastrointestinal tract, unspecified	1.00

To evaluate the performance of the proposed method, we compared the classical SVM-RFE to SVM-RRFE (referred to as RRFE in this paper) proposed by Johannes et al. [8]. The original RRFE integrates GeneRank algorithm into SVM-RFE feature selection approach to reduce the instability of selected features by incorporating prior knowledge. The SVM in combination with RFE was introduced for gene selection in bioinformatics by Guyon et al. [6]. The SVM-RFE feature selection method is based on linear SVM used for feature ranking. Each feature is ranked by its impact on the weight vector, based on the Lagrange multipliers in the SVM optimization problem [14]. In the final step of each iteration, all the genes are ranked and a pre-selected number of the lowest-ranked genes is eliminated. By default, a single gene is eliminated in each round. However, it is more common to eliminate a certain percentage of features per iteration in high-dimensional settings.

Not only high dimensionality, but also an extremely high number of samples can be met in hospital discharge data. Therefore, it is

usually not possible to use all samples when building the classification model. One solution of this problem is the inclusion of prior knowledge in the form of disease networks, as described in Section 2.3. RRFE uses GeneRank algorithm to obtain additional feature weighting information based on network and learning data and integrates it directly in the SVM. The new feature weights are computed as

$$\phi(w_j, r_j) = \frac{w_j}{\rho(r_j)}$$

combining feature weights from SVM (w_j) and rank of GeneRank weight $\rho(r_j)$. This way the feature ranking considers the impact of the SVM weights and the connectivity of a feature in the prior knowledge network. Due to computational reasons the misclassification penalty parameter *C* in SVM was set to 0.1 in all experiments described in this paper. However, there are some important differences between the application of the original RRFE and our application to hospital discharge data that are described in the following section.

To cope with the problem of high imbalance in the target class, we use repeated random subsampling with target class balancing. In each iteration of classification performance evaluation, we randomly select 10,000 samples, where the distribution of classes is not manipulated. This set of instances is used for testing, as it resembles the original class distribution. In the second step of each evaluation cycle, we randomly sample p_rN positive and N-p_rN negative samples, where p_r represents a ratio of positive samples and N represents a number of all samples used for training the classification model. In the first experiment, we used equally balanced classes (p_r=*0.5*) trained using 1000 samples. According to a study by Moturu et al. [11], where a similar dataset was used, the optimal classification performance should be achieved with p_r=*0.75*, and we therefore conducted a second experiment where 75% of samples in the training set were sampled from all available target class samples. Each random subsampling evaluation was repeated 10 times for all target diagnosis codes.

As already mentioned in Section 2.3, we used relative risk (RR) based networks in all experiments. To avoid bias, we built the disease network on a dataset containing hospital discharge records for 2008, while all classification evaluations were done on data from 2009. A sparse matrix with 14,315 ICD9-CM diagnosis codes and 8,921,946 interactions among them was constructed from 2008 data. To test the impact of data availability in the network construction process, we constructed another disease network where we used all available data from 2000-2008 NIS datasets. Altogether, we used 58,761,912 records to construct a disease network containing 18,288,394 relative risk values for different pairs of diagnoses. All experiments were implemented in R [13], using pathClass package [8].

4. RESULTS

4.1 Classification

In the first experiment we compared classical SVM-RFE to the RRFE with 50% and 75% of positive class samples used in random subsampling step. The network for the RRFE method was constructed from all available samples for 2008. As in [8] we use an elimination rate of 10% in each iteration of RFE. The results are displayed in Table 2. One can observe minimal differences between two compared methods. In case of Hyperlipidemia (272.4) the best performance shifted from SVM-RFE to RRFE, while in case of Iron deficiency anemia (280.9) the shift in the opposite direction occurred. Overall, based on our experiments, we cannot confirm the improvement of the results when the percentage of target class is increased as in [11].

Table 2. Comparison of AUC for RRFE and SVM-RFE with 10% removal rate.

Positive Ratio Disease Code	RRFE	SVM-RFE
0.5 (average)	**0.726 ± 0.068**	0.723 ± 0.071
272.4	0.745 ± 0.014	**0.757 ± 0.010**
278.00	**0.723 ± 0.024**	0.706 ± 0.020
280.9	**0.668 ± 0.021**	0.667 ± 0.023
285.9	**0.657 ± 0.021**	0.649 ± 0.014
578.9	0.837 ± 0.019	**0.838 ± 0.011**
0.75 (average)	**0.728 ± 0.072**	0.726 ± 0.074
272.4	**0.767 ± 0.006**	0.765 ± 0.005
278.00	**0.707 ± 0.016**	0.690 ± 0.017
280.9	0.670 ± 0.020	**0.676 ± 0.025**
285.9	**0.652 ± 0.024**	0.651 ± 0.021
578.9	0.843 ± 0.010	**0.846 ± 0.019**

In the second experiment, we repeated the first experiment with the only difference being in the rate of feature removal. Here, we removed 50% of features in each iteration of RFE. The results displayed in Table 3 confirm our expectations that RRFE might prove more effective when larger sets of features are removed. This could be due to the fact that SVM-RFE is a very unstable feature selection method which intensifies when the removal rate is increased [18].

Table 3. Comparison of AUC for RRFE and SVM-RFE with 50% removal rate.

Positive Ratio Disease Code	RRFE	SVM-RFE
0.5 (average)	**0.743 ± 0.067**	0.732 ± 0.069
272.4	**0.769 ± 0.008**	0.761 ± 0.009
278.00	**0.731 ± 0.016**	0.718 ± 0.019
280.9	**0.694 ± 0.033**	0.678 ± 0.020
285.9	**0.668 ± 0.016**	0.659 ± 0.009
578.9	**0.850 ± 0.016**	0.844 ± 0.018
0.75 (average)	**0.739 ± 0.072**	0.719 ± 0.082
272.4	**0.777 ± 0.006**	0.765 ± 0.011
278.00	**0.717 ± 0.014**	0.698 ± 0.028
280.9	**0.687 ± 0.029**	0.648 ± 0.040
285.9	**0.662 ± 0.017**	0.640 ± 0.020
578.9	**0.855 ± 0.020**	0.844 ± 0.019

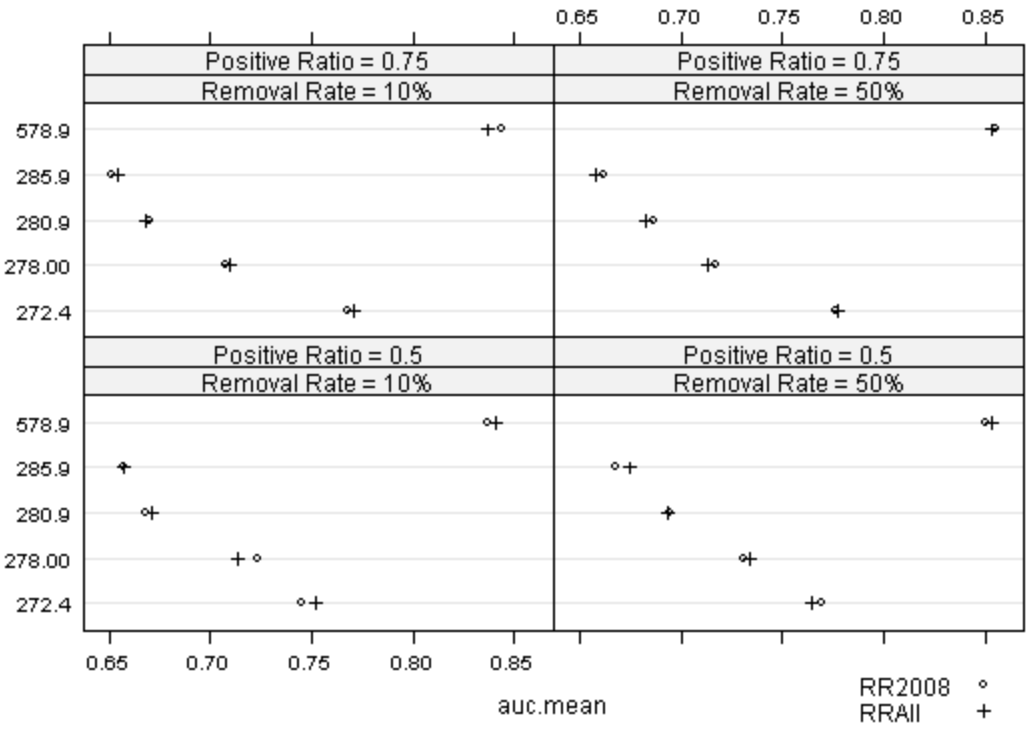

Figure 2. Comparison of AUC for RRFE and SVM-RFE with 10% removal rate (subsampling with 75% positive class).

In the case of hospital discharge classification problems, it is extremely important if we can afford a less complex and faster method, which is the case in the 50% removal rate example. Comparing the results from both experiments, one can notice that the less complex 50% removal approach actually improved the performance of the classification in almost all cases.

To further analyze the performance of the compared methods, we conducted Wilcoxon's signed-rank test and compared results from both experiments for 5 disease classification problems together (*n=50*). In the case of a 10% removal rate, the difference in AUC is not statistically significant ($p = 0.060$). On the other hand, the difference in AUCs between SVM-RFE and RRFE for 50% removal rate is statistically significant ($p < 0.001$).

In our final experiment, we wanted to test how the availability of data influences the effectiveness of the disease network-based feature selection approach in RRFE. Therefore, we constructed an additional network from 9 years of data (2000-2008) and ran the first two experiments for RRFE again. To our knowledge, this is the largest disease network created to date.

However, our results demonstrate that a larger network does not produce significantly better results in classification performance (Figure 2). Wilcoxon's signed-ranks test confirms our observations from Figure 2 with *p=0.5867*. Based on the results of this experiment, we can conclude that using more recent and less complex disease networks does not significantly impact the classification performance. One of the reasons for weak performance of the large network may also lay in the fact that some disease codes change from year to year. Therefore, it does not make sense to store disease interactions that no longer exist in the newer versions of ICD9-CM coding.

4.2 Stability

To compare the stability of the selected features over multiple random subsamples and different experimental settings for both SVM-RFE and RRFE, we measured the frequency of occurrences when a specific disease code was included in an optimal set of features (Table 4). The optimal set of features is a feature set obtained after all RFE iterations have been executed and a set with the highest performance is chosen. Altogether we compare optimal sets from different experimental settings done during the comparison of performance, described in Section 4.1.

Table 4. Frequency of disease code selection in the optimal feature sets for Hyperlipidemia (272.4) classification.

SVM-RFE		RRFE	
ICD9-CM Code	**Freq.**	**ICD9-CM Code**	**Freq.**
401.9	80	250.00	80
414.01	78	401.9	80
V27.0	78	414.01	80
250.00	70	V27.0	80
272.0	70	403.90	79
403.90	64	530.81	78
401.1	44	244.9	75
434.91	44	272.0	74
477.9	44	428.0	74
244.9	42	V45.82	74

Table 4 compares consistency of selection for the most frequently selected codes in the first classification problem – i.e. Hyperlipidemia classification. The compared optimal sets are very similar in their mean number of features in an optimal set (151.85 for SVM-RFE and 151.26 for RRFE). As expected, the consistency of selected features in RRFE selection is much higher as compared to SVM-RFE. There is only one disease code (401.9 - Unspecified essential hypertension) that was chosen in all experiments by both feature selection methods.

On the other hand, there are some codes that were not represented in a significant number of SVM-RFE based optimal feature sets. For example, diagnosis code 250.00 (Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled), was missing in 10 out of 80 optimal sets when SVM-RFE was used. RRFE included this feature in all random subsampling datasets with different experimental settings.

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

This paper presents an adaptation of the RRFE method for feature selection, originally used in bioinformatics, as well as application of this method to feature selection in imbalanced high-dimensional hospital discharge data. The results confirm the advantages of the method that were successfully used in the bioinformatics domain by demonstrating the increased stability of selected features. Furthermore, we observe significant improvements of classification performance when large batches of features are eliminated. Due to an extremely large number of samples in hospital discharge datasets, this is especially important. In our case, a dataset consisting of nearly 7 million samples collected in one year was used for learning. Since this dataset represents hospital discharge data for approximately 20% of discharges from U.S. hospitals, we can estimate the number of hospitalizations approaches 100,000 per day. In such large data settings it is important to effectively select the important features for classification.

Although we evaluated the classification performance of the proposed feature selection solution using SVM classifier, it would be possible to use it in combination with another, preferably simpler, classification model. Use of a simple classifier makes the proposed method more appropriate for scalability to large data settings.

6. ACKNOWLEDGMENTS

This research was supported by the Slovenian Research Agency through a bilateral project grant ARRS-BI-US-JR/2011/16.

7. REFERENCES

[1] HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). 2000-2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/nisoverview.jsp

[2] Chawla N.V., Japkowicz N. and Kotcz A. 2004. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations Newsletter, 6.

[3] Cios, K. J. and Moore, G. W. 2002. Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1-2), 1-24.

[4] Cun, Y. and Frohlich, H. 2012. Biomarker gene signature discovery integrating network knowledge, Biology, 1, 5-17; DOI=http://dx.doi.org/10.3390/biology1010005.

[5] Davis D., Chawla N., Christakis N. and Barabási A. 2009. Time to CARE: a collaborative engine for practical disease prediction. Data Mining and Knowledge Discovery. pp 388–415.

[6] Guyon I., Weston J., Barnhill S. and Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389-422.

[7] Hidalgo C.A., Blumm N., Barabási A. and Christakis N.A. 2009. A dynamic network approach for the study of human phenotypes. PLoS Computational Biology, 5.

[8] Johannes M. Brase J., Frohlich H. Sultmann H., Beissbarth T. 2010. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics, 26, 2136–2144.

[9] Khalilia M., Chakraborty S., Popescu M. 2011. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making, 11:51.

[10] Morrison J.L., Breitling R., Higham D.J. and Gilbert D.R. 2005. GeneRank: using search engine technology for the analysis of microarray experiments. BMC Bioinformatics, 6, DOI=http://dx.doi.org/10.1186/1471-2105-6-233.

[11] Moturu S.T., Liu H. and Johnson W.G. 2008. Understanding the effects of sampling on healthcare risk modeling for the prediction of future high-cost patients., in Ana L. N. Fred; Joaquim Filipe & Hugo Gamboa, ed., 'BIOSTEC (Selected Papers)' , Springer, pp. 493-506.

[12] Page L. Brin S. Motwani R. and Winograd T. 1999. The PageRank citation ranking: bringing order to the web; Technical Report 1999-66; Stanford InfoLab: Stanford, CA, USA.

[13] R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[14] Scholkopf, B. and Smola, A. 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA.

[15] Siebourg J., Merdes, G., Misselwitz B., Hardt W.D. and Beerenwinkel N. 2012. Stability of gene rankings from RNAi screens. Bioinformatics, DOI=http://dx.doi.org/10.1093/bioinformatics/bts192.

[16] Steinhaeuser K. and Chawla N.V. 2009. A network-based approach to understanding and predicting diseases. Social Computing, Behavioral Modeling, and Prediction, Springer, 209-216.

[17] Stiglic G. and Kokol P. 2010. Stability of ranked gene lists in large microarray analysis studies. Journal of biomedicine & biotechnology, 616358.

[18] Tang Y., Zhang Y.-Q. and Huang Z. 2007. Development of two-stage SVMRFE gene selection strategy for microarray expression data analysis, IEEE Trans. Comput. Biol. Bioinformatics, vol. 4, no. 3, pp. 365–381.

[19] Tang Y., Zhang Y.Q., Chawla N. 2009. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B, Cybern 39(1):281–288.

[20] Turan N., Ghalwash M.F., Katari S., Coutifaris C., Obradovic Z. and Sapienza C. 2012. DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? BMC Medical Genomics, 5:10.

[21] Xing E., Jordan M. and Karp R. 2001. Feature selection for high-dimensional genomic microarray data. Proc. 15th International Conference on Machine Learning, pp. 601-608