Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 13-020

Correlation based Feature Selection using Rank aggregation for an Improved Prediction of Potentially Preventable Events

Chandrima Sarkar, Prasanna Desikan, and Jaideep Srivastava

June 12, 2013

Correlation based Feature Selection using Rank aggregation for an Improved Prediction of Potentially Preventable Events

Chandrima Sarkar[1,2] MS, Prasanna Desikan[2] PhD, Jaideep Srivastava[1] PhD
[1]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN; [2]Allina Health, Minneapolis, MN

Abstract

This paper presents a methodology for developing a novel feature selection model that will help in a more accurate and robust prediction of patients with the risk of Potentially Preventable Events (PPEs). PPEs are admissions, readmissions, complications and emergency department visits that could have been avoided if the patient had been given the appropriate interventions. Various clinical factors and patient health conditions can affect a patient's chance of developing the risk of PPE. We propose a robust Correlation based feature selection method using Rank Aggregation (CRA) which helps to identify the key contributing factors for the prediction of PPE. Unlike existing feature selection techniques that causes bias by using distinct statistical properties of data for feature evaluation, CRA uses rank aggregation thus reducing this bias. The result indicates that the proposed technique is more robust across a wide range of classifiers and has higher accuracy than other traditional methods.

Introduction

As the healthcare industry moves from a fee-for-service model to pay-for-performance model, there is large emphasis on managing the health of the populations as whole. Government and private payers are designing their reimbursement schemes based on the performance of the health of the population. This drives provider organizations to be more efficient in providing care and reducing costs [1], [13]. A potential solution to this wastage reduction for improved care is the detection and prevention of potentially preventable events. Potentially Preventable Events (PPEs) are hospital visits including visits to the admissions, readmissions, and emergency department (ED) that could have been avoided if the patient had been given and complied with appropriate interventions in the ambulatory setting. Proper PPE detection can assist greatly in cost reduction as well as can serve as an early indicator of an impending health problem, thereby saving a patient from future hospital visits. Though detection of PPE is a very challenging task, data mining and machine learning techniques using Electronic Health Records (EHR) has proved to be very efficient in successfully building clinical decision support systems that satisfactorily predicts the onset of PPEs in a large population of patients [3]. One of the most important challenges in finding significant statistical patterns using machine learning techniques is due to the high dimensional nature of the EHR data. In general healthcare data contains more than 1000 variables and is highly sparse which makes statistical and empirical analysis complex and cumbersome in inferring a prediction outcome. On top of that, one always faces the dilemma of choosing the right statistical measure for evaluating an outcome. Apart from analysis and improvement of PPE prediction outcomes, non-redundant and relevant variables selection is an important factor PPE analysis with EHR data. Variables selection techniques for ultra-high dimensional data with over 1000 dimensions are a challenging task due to the computational complexity, statistical accuracy, and robustness (consistent accuracy) across different types of classifiers. Existing feature selection techniques exploit distinct statistical properties of data (e.g. chi square statistics, information gain) for feature evaluation. This causes a bias towards the specific statistical properties that they use. In this paper, we propose a novel Correlation based feature selection method using Rank Aggregation (CRA) which helps in alleviating the bias introduced by the measures of statistical dispersion in large scale very-high dimensional data. We also define an evaluation measure of CRA for robustness namely Robustness Index who's algorithmic description has been provided in the consequent sections. We show that our method is more robust over a wide range of classifiers i.e. the classification errors generated using features selected from CRA is consistently lower than most of the contemporary feature selection techniques. We evaluate robustness of CRA based on mean classification error and the consistency of the accuracy across multiple classifiers, in order to help compare CRA with other contemporary methods. We use our technique to predict PPE with data from a large healthcare organization. We chose this data set to test the ability of this novel data mining based approach to identify relevant variables and improve prediction accuracy because of the complexity of PPE and the very high dimensional nature of the data with a large number of clinical variables. To the best of our knowledge, such approach has never been used in this domain to improve upon accuracy for efficient PPE prediction.

To summarize, this paper makes the following contributions:

- Development of a novel correlation based feature selection technique designed to obtain the most significant variables for prediction of PPE.
- Improving the overall prediction accuracy of PPE and robustness across a variety of classifiers using very high dimensional healthcare (EHR) data.

The remainder of this paper is organized in the following order - *Background and Related Work* Section describes a brief detail of the background and motivation of this research and the related work in this area, *Method* section describes the methodology in detail, In *Results* section, we analyze and discuss the results obtained through various empirical analysis and evaluations, *Discussion and Conclusion* section, *Acknowledgement* section and *References*.

It should be noted that in this paper the term 'features' has been referred as 'variables' or 'attributes' based on the context of the discussion. However, all three terms have the same intended meaning.

Background and Related Work

Feature selection is a key step in machine learning for reducing dimensionality, increasing learning accuracy, and improving result comprehensibility. It has immense potential to enhance data mining in the healthcare domain as shown in previously studies in various important healthcare applications such as surveillance and infection control, work-flow optimization, distributed medical databases and treatment support for patients with heart disease [3], [11], [12], [16], [18], [8]. However, the recent increase of dimensionality of EHR data (which extends over 1000 variables and over 100,000 instances), has posed a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. This enormity poses challenges to many machine learning algorithms with respect to scalability and prediction performance. The reason is that the data sets with thousands of features features/variables can contain large amount of information that does not contribute to the predicting ability and also may greatly degrade the performance of classification algorithms. This makes feature selection an extremely important preliminary step to any machine learning tasks when dealing with very high dimensional data such as EHR. There are mainly two types of feature selection mechanisms known as Filter approach and Wrapper approach. Filter approaches use an evaluation function that relies only on properties of the data. Wrapper approaches use the inductive or learning algorithms to estimate the value of a given subset. In the former, the measure of significance or relevance is defined independently from the learning algorithm while in the latter, the measure of significance is directly defined from the learning algorithm, for example, cost is one of the learning ability of the model. In this paper we focus on ranker based filter technique since there is more advantage in using filter approach as compared to that of the wrapper approach [10]. Filter method is fast and simple which makes it more suitable for high dimensional data [19] than Wrapper methods because when the dimensionality becomes very large, filter method has lesser computational time complexity.

There is extensive body of existing research in filter feature selection. Different feature selection algorithms use different statistical properties of data such as algorithms based on correlation measure between features/ variables [19], algorithm known as Relief, which randomly samples a given number of instances from the training set and updates the relevance estimation of each feature based on the difference between the selected instance and the two nearest instances of the same and opposite classes [9]. But different techniques have their individual drawbacks such as Relief does not take into account the redundancy in the features. Features highly correlated to each other still get selected and past researches show along with irrelevant features, redundant features also affect the speed and accuracy of classification and prediction. This becomes more important when the dimensionality is very high and the chances of having redundant and irrelevant features increases. On a broader scale , this shows that different feature selection techniques uses different statistical properties which can cause bias towards achieving the most optimal solution for a classification problem given a very high dimensional data set. Attempts has been made in the past for reducing this statistical bias by using rank aggregation on different feature selection techniques use that uses different statistical properties such as information gain, gain ratio chi square statistics , mutual information [14] which has shown that ensemble feature selection improves the result considerably. But the aforementioned work has shown this to be true for high dimensional data with less than 200 features/ variables. Ensemble feature selection has also been used in ultra-high dimensional genomic data and has shown to yield notable improvement in the result. But EHR or healthcare data has different statistical properties as that of data with lower dimension or very small sample size as that of the genomic data [17]. These works has used simple average method for performing rank aggregation. Moreover, to the best of our knowledge, all the past works in ensemble based feature selection

approaches has never considered data redundancy and relevancy while aggregating based on some statistical properties. In our paper we extend the idea of rank aggregation mechanism based on correlation among the features that reduces redundant and irrelevant data while considering important statistical property. This unique combination of correlation and other statistical properties of data such as information gain, chi square statistics and mutual information make our approach least biased towards any statistical property. We show in this paper that our method is more robust over a wide range of classifiers i.e. the classification errors generated using features selected from CRA is consistently lower than most of the contemporary feature selection techniques.

Methods

Preliminaries

In this paper we propose a feature selection approach based on Correlation and Rank Aggregation (CRA) to find significant variables from high dimensional healthcare data. We evaluate our feature selection method by performing classification on the extracted features to verify the prediction accuracy. We evaluate robustness of our approach by comparing the variance in error rate obtained from classification using our feature selection technique with other feature selection techniques such as information gain, symmetric uncertainty and Chi-square.

The rationale behind using correlation and rank aggregation based feature selection is primarily, to eradicate redundant and irrelevant features with a positive effect on the predicted outcomes and secondly, to capture important variables which may prove as essential factors contributing to a successful outcome of PPE. The novelty of our approach is the use correlation as a filtering and feature-reduction step which then feeds into rank aggregation techniques for feature selection. Our feature selection technique CRA yields a list of significant variables that can be used to select the key contributing factors of patients with the risk of PPE. CRA is unique because it helps to reduce biases caused by different feature selection techniques with distinct statistical properties of data while providing higher accuracy, sensitivity, and specificity, which are often very hard to achieve with single statistical models especially with very high dimensional healthcare data containing over 1000 variables. After we obtain the final list of significant features we used this feature set for prediction of PPE using a wide range of classification algorithms. Some of the classifiers that we used were - Naive Bayes, J48, K nearest neighbor, AdaBoostM1, Logistic Regression, Bagging and ADTree. The result showed that the accuracy of our approach is higher than the other feature selection methods such as information gain, symmetric uncertainty and Chi-square. The diagram of the entire procedure is shown in Figure 1.

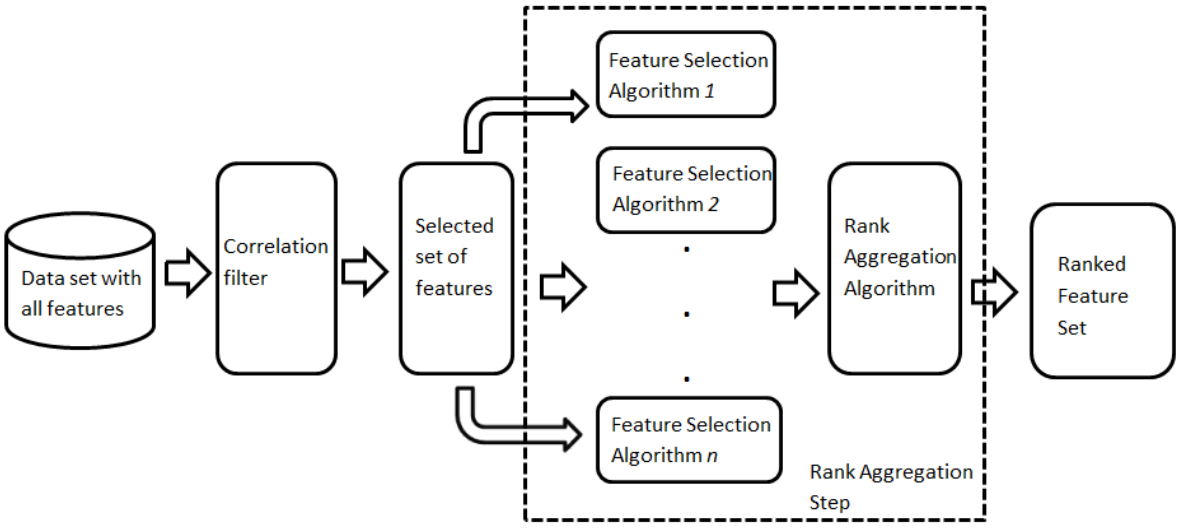

Figure 1: Flow diagram of the entire approach. It consists of two main steps: correlation filter step and rank aggregation step.

Detail

Feature Selection: Feature Selection is one of the most important pre-processing tasks in data mining. The most important reason is that not all features/attributes are relevant to a given problem. Some of the features interfere and

reduces prediction accuracy. In addition, dealing with larger number of features is most of the time costly. Therefore reducing dimensionality is rather important when dealing with very high dimensional data such as in Electronic health records. Feature selection also helps in providing a better understanding of the process that generated the data [5]. In this paper we use a novel technique which utilizes two main properties of dimensionality reduction namely, correlation among the features and rank aggregation based on different statistical properties of data [14]. There are two main steps in our entire feature selection process namely Correlation Filter and Rank Aggregation. We discuss about each of these steps in detail below.

1) Correlation Filter: The correlation filter step is based on the concept of reduction of redundant and irrelevant variables [19]. This step aims in identifying relevant features from the huge feature set and reducing redundancy among the relevant feature if it exists with the help of correlation measure. The guiding principle is that a feature is not redundant if it is relevant to the class variable but is not redundant to any of the other relevant features. Hence, this correlation filter weeds out all those features which are either not highly correlated with the response variable or has low correlation among the other features. The end result of this step is a feature set which has variables that are highly correlated with the response variable but has less correlation with the rest of the features. This correlation filter step used has been implemented based on CFS (Correlation feature selection algorithm) [7], [6]. The output of this step is a feature subset with relevant and non-redundant features. This output is then utilized as the input to the next rank aggregation step.

2) Rank Aggregation: Rank aggregation is the procedure that takes as input multiple rankings of a fixed set of alternatives (or candidates), and aggregates them into a single consensus ranking of the candidates. Rank aggregation can be done in various ways namely Borda and Kemeny [2], [4]. We use rank aggregation mechanism based on Borda [2], [15]. For rank aggregation we use a position based scoring mechanism to calculate the final score of a feature. We then sort each feature based on their final score and rank them from highest (being rank 1) to lowest (being rank n, for n features) according to their final scores generated as given in equation 1.

$$\boldsymbol{score_{final}} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} \boldsymbol{score_{pos(i)}} \qquad (1)$$

Where n is the number of features selection techniques used. $pos(i)$ is the i^{th} position of a feature ranked by the ranker n (which in our case is the different feature selection techniques based on information gain, symmetric uncertainty and chi square statistics), obtained from feature's i^{th} position. $Score_{final}$ is the final score obtained after rank aggregation. We used the following feature selection techniques for using the ranks generated by each of them according to their distinct statistical properties of data

- Information gain Attribute Evaluation This algorithm evaluates features individually by measuring their information gain with respect to the classes.
- Symmetrical Uncertainty Attribute Evaluation This algorithm evaluates features individually by measuring their symmetrical uncertainty with respect to the class.
- Chi-Square Statistics This algorithm evaluates features individually by measuring their chi-squared statistic with respect to the classes.

The main benefit of using correlation based feature selection technique using rank aggregation is the fact that features generated are not only relevant and non-redundant but also are less biased towards any specific statistical properties of data such as information gain or mutual information. The ensemble nature of rank aggregation helps to make it significant with respect to the different statistical properties of data. Another interesting point we note was that our feature selection technique generated feature set which was robust with consistent accuracy across a wide variety of classification algorithms. Moreover, results were obtained which showed that prediction accuracy of PPE increased with a decrease in the mean error across different classifiers. We discuss the result in detail in our Results section.

A. Data set and technology used

The primary data source was EHR data (from a large health system in the metro region) for the years 2010 and 2011 for patients with clinic visits for ambulatory care sensitive conditions. These data consists of over 1000 variables and about 100000 patient records focused on clinic visits for ambulatory care sensitive conditions. Data was extracted from EHR data from the Enterprise Data Warehouse. Data extracted comprised of demographic

information, lab results, flow sheets, problem lists and clinic visits for the patients. Weka libraries and JAVA were used for implementing the feature selection technique.
1) Preprocessing of the data: This data was pre-processed to handle missing values and aggregate other pieces of information to prepare for data mining tasks. Matlab was used for data pre-processing. Response variable used for prediction is PPE. PPE is a binary class with value - YES (detected with a risk of PPE) or NO (not detected with a risk of PPE).

2) Evaluation measure: In order to evaluate the performance of our feature selection technique, we compared our technique with other rank aggregation algorithms using distinct statistical properties of data such as information gain, symmetric uncertainty and Chi square statistics, based on two evaluation measures namely –

- Accuracy of PPE prediction - accuracy was calculated as the total percentage of correctly classified instances by a given set of classifiers. The final ranked feature set obtained was sampled into 30 feature subsets in ranked order of decreasing significance (top 1, top 2 ... top 30). Figure 2 shows results of accuracy of PPE prediction for top 4, top 9, top 14 and top 19 (these numbers were chosen randomly for the purpose of examples in this paper). Accuracy was measured until top 30 which was heuristically determined for evaluation purpose as well as to keep the dimension to a lower order. Accuracy of our feature selection technique was compared against all other feature selection methods mentioned previously in this paper.

- Mean classification error - this is calculated as the mean classification error in prediction of PPE from a given set of classifiers. Feature set generated from CRA was used for calculating mean classification error from a given set of classifiers. This is compared against the mean error calculated from the other feature selection techniques using information gain, symmetric uncertainty and chi square statistics.

- **Robustness Index** - We define robustness index as the measure that evaluates the consistency in classification accuracy for PPE prediction across multiple classifiers. Intuitively, the consistency depends on whether the classification accuracy (or error) increases or decreases as compared with respect to other methods. The robustness index is measured according to the following algorithm –

    - Input – Classification models M_i , where $1 \leq i \leq n$, n is the number of classifiers used; Feature set f_k generated from p feature selection techniques, where k is number of top features

    - Output – Robustness Index for each feature selection method

    STEP 1: For each Mi , $1 \leq i \leq n$
    a. Determining the classification error with f_k feature subset.
    b. Rank each p according to the classification error obtained using f_k , for e.g. – the p producing the lowest error will be given the highest rank (1^{st} position), p with the second highest error will be given the 2^{nd} position/rank and so on
    STEP 2: Aggregate the ranks for each M_i , for each p according to equation 1.
    STEP 3: Analyze the aggregated ranks and denote them as robustness index where lower the rank of a p , higher is its robustness index (e.g. p with rank 1 has the highest robustness index).

Thus, more a feature selection technique has a consistent lower classification error across different classification model, higher is its robustness Index. It also denotes that it has consistent higher classification accuracy across different classifier. A system that produces consistent outcomes is more advantageous. This is because; it is not an easy task to guess the best classification model to use prior to actually using that model. Our results show that prediction of PPE using features produced by CRA is consistently high across various classifiers. A robust technique helps one to choosing classification model with the minimum risk of choosing an inappropriate model.

3) Classification Algorithms used: Classification algorithm used for prediction of PPE are namely Naive Bayes, J48, 3 nearest neighbor, AdaBoostM1, Bagging and AdTree. We evaluated our method based on a variety of classification algorithms in order to demonstrate the robustness of our approach. Our feature selection technique can

be used along with a variety of prediction measures including rule based, Bayesian Network. Classification tree, ensemble based and even statistical measure like regression due to robust characteristics of this approach. Results from our analysis shows that we get consistently higher prediction accuracy and robustness across all different classification techniques considered in this

Results

We focus our analysis on the prediction accuracy obtained with these key factors using different classification algorithms. We do not discuss the actual key contributing factors identified / the significant variables detected by our technique. This analysis helps us in determining the significance of the identified variables. For this purpose we conducted a comparative analysis based evaluation of our feature selection method in order to analyze its performance with respect to accuracy of PPE prediction, average f-measure from prediction of PPE, mean classification error in prediction across a variety of given classifiers such a Naive Bayes, J48, 3-nearest neighbor, AdaBoostM1, bagging and AdTree and the robustness of our algorithm. We analyzed and evaluated the output of our rank aggregation algorithm by comparing the results against the other single feature selection algorithms such as Information gain Attribute Evaluation, Symmetric Uncertainty Attribute Evaluation and Chi square Attribute Evaluation.

The results of accuracy are given in Figure 2A, 2B, 2C and 2D. These results show the comparative analysis of prediction accuracy of PPE using feature set generated from our CRA method with three other feature selection methods mentioned in the previous paragraph. The reason for not choosing feature selection methods such as PCA or SVM is that they scale up the original feature to a different dimension which most of the time is difficult to analyze actual clinical variables in a case especially like PPE since we aim in identifying the key contributing factors to PPE prediction. For this comparative analysis we have randomly selected top *n* number of features from our final ranked list to generate the final feature set. We show in Figure 2A, 2B, 2C and 2D, four consecutive feature subsets namely - top 4, top 9, top 14 and top 19. In Figure 2A we have chosen top 4 factors or variables / features for predicting PPE. We can see that the accuracy obtained is higher than the other three methods in almost all the cases. Similarly in Figure 2B where we use top 9 features J48 and Bagging gives the best accuracies and Naïve Bayes the worst accuracy. But in almost all of the classifiers the accuracy by CRA is consistently higher than other methods. This is true for results obtained using top 14 features shown in Figure 2C and top 19 features shown in Figure 2D. Though we found that for this data set Symmetric Uncertainty gives a comparative accuracy, but our method has a better accuracy generated from almost all of the top n features using different classifiers. We performed this analysis over top 30 features and found similar result having CRA obtaining higher accuracy in most of the cases over any other methods than other traditional methods. However, we randomly picked top 4, 9, 14, 19 features for analysis in this paper due to page limitations.

In Table 3 we show the results of comparison of robustness index for CRA with the other traditional feature selection techniques. We calculate the robustness index as described in the Algorithm in the *Methods* section.
Since the robustness index is calculated from rank of a technique based on how it performs (accuracy) with respect to the other methods, it is clear that lower the (rank) i.e. the value of index, more robust is the technique i.e. Robustness index of 1as compared to 3 means the former is more robust than the latter. We can see in the table that CRA scores over all the other methods in all the cases (different feature subsets). CRA has robustness index of 1 in all the cases. This shows us that CRA is a more robust technique than the other traditional feature selection techniques. The intuition behind this analysis is that, when one is unable to decide on the best classification algorithm to use, CRA will guarantee to give a lower classification error with most of the classifiers.

In Table 4 we compare the average f-measures (defined as the harmonic means of precision and recall) from the two classes, generated using CRA with other feature selection methods using different classifiers. We take top 4, top 9, top 14 and top 19 feature subsets like the previous case to calculate f-measures using different classifiers such as Naive Bayes, J48, 3-nearest neighbor, AdaBoostM1, bagging and AdTree. We find that even CRA have a higher average f-measure than the other three methods in almost all of the cases f-measures generated from 4 different feature subsets). This shows CRA gives a better sensitivity and specificity with different classifiers.

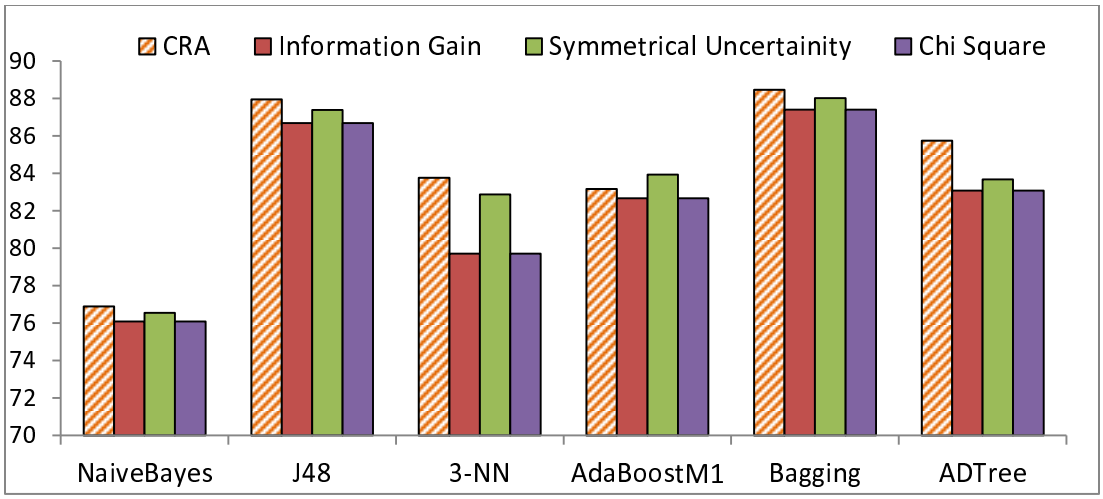

Figure 2A : Comparison of classification accuracy for (Top 4 features). X –axis Classifiers; Y axis – Classification Accuracy

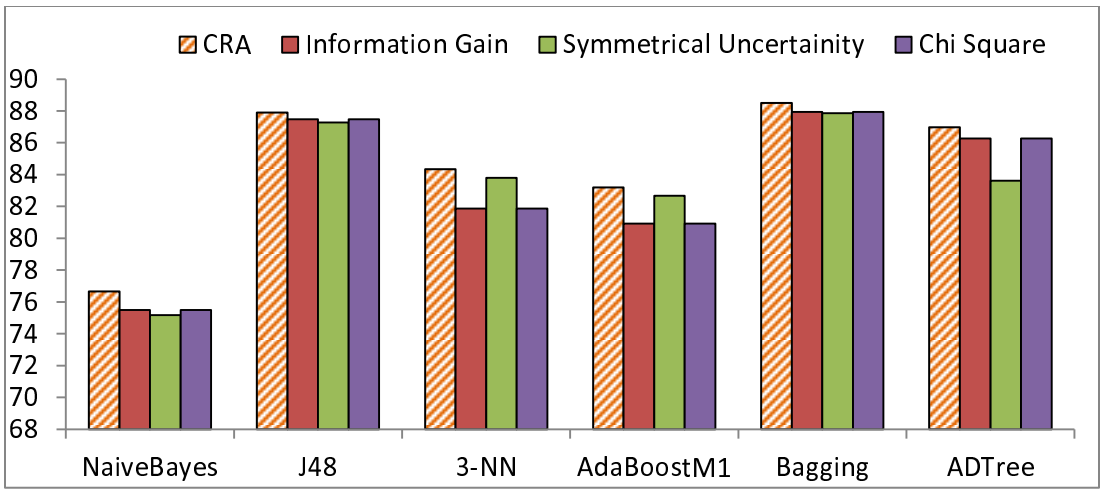

Figure 2B : Comparison of classification accuracy for (Top 9 features). X –axis –Classifiers used; Y axis – Classification Accuracy

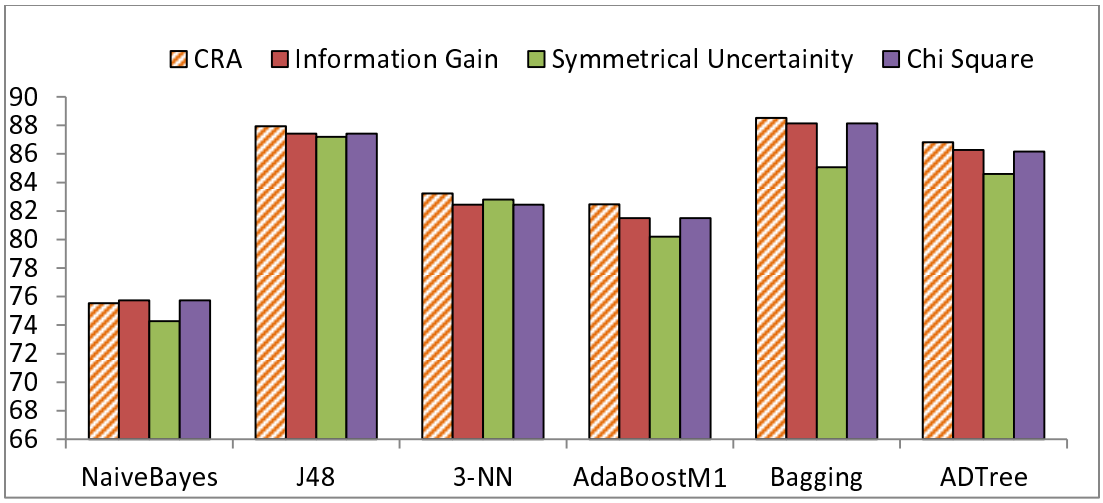

Figure 2C : Comparison of classification accuracy for (Top 14 features). X –axis –Classifiers used; Y axis – Classification Accuracy

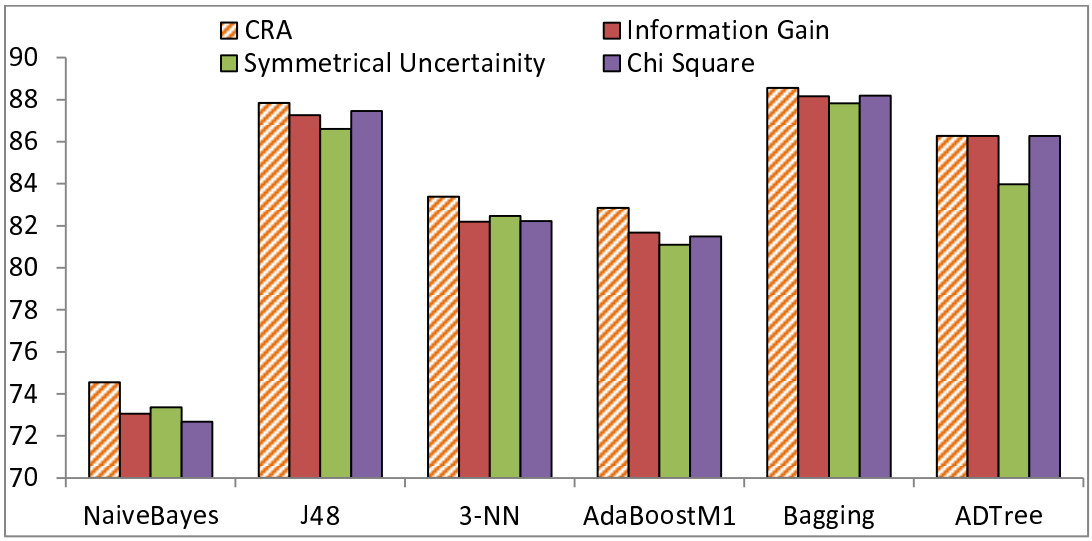

Figure 2D : Comparison of classification accuracy for (Top 19 features). X –axis –Classifiers used; Y axis – Classification Accuracy

	Top *N* features	CRA	Information gain	Symmetrical Uncertainity	ChiSquare
Robustness Index	4	1.00	3.00	2.00	3.00
	9	1.00	3.00	2.00	3.00
	14	1.00	3.00	2.00	3.17
	19	1.00	2.67	3.00	2.50

Table 3: Robustness index for all feature selection methods using Naive Bayes, J48, 3-nearest neighbor, AdaBoostM1, bagging and Adtree. Lower the value of index, more robust is the technique i.e. Robustness index of 1as compared to 3 means the former is more robust than the latter.

	Top 4						Top 9					
	Naive Bayes	J48	Nearest Neighbor	AdaBoost	Bagging	ADTree	Naive Bayes	J48	Nearest Neighbor	AdaBoost	Bagging	ADTree
CRA	0.55	0.76	0.65	0.45	0.77	0.72	0.57	0.77	0.69	0.62	0.78	0.74
Information gain	0.52	0.73	0.70	0.64	0.77	0.67	0.56	0.76	0.67	0.53	0.77	0.74
Symmetrical Uncertainity	0.48	0.67	0.68	0.64	0.69	0.66	0.55	0.74	0.69	0.60	0.77	0.64
ChiSquare	0.52	0.73	0.70	0.64	0.77	0.67	0.56	0.76	0.67	0.53	0.77	0.74
	Top 14						Top 19					
	Naive Bayes	J48	Nearest Neighbor	AdaBoost	Bagging	ADTree	Naive Bayes	J48	Nearest Neighbor	AdaBoost	Bagging	ADTree
CRA	0.57	0.77	0.69	0.62	0.78	0.74	0.58	0.76	0.69	0.63	0.78	0.74
Information gain	0.57	0.77	0.67	0.57	0.78	0.73	0.57	0.76	0.67	0.59	0.77	0.73
Symmetrical Uncertainity	0.56	0.75	0.67	0.50	0.77	0.67	0.56	0.75	0.67	0.54	0.77	0.68
ChiSquare	0.57	0.77	0.67	0.57	0.78	0.73	0.57	0.77	0.67	0.57	0.77	0.73

Table 4: F measure comparison between CRA and other traditional feature selection algorithms

In Table 5 we show the result of mean classification error generated using different classifiers using feature set obtained from CRA and other 3 methods. In this case, less is the mean error; more is the robustness of the feature

selection method. We see that using CRA method the mean classification error generated using different classifiers is least (for almost all of the cases). This proves that CRA method is not only robust but also highly robust across all the classifiers.

Top *N* Features	Mean Classification error in %			
	CRA	Information gain	Symmetrical Uncertainity	ChiSquare
4	15.67	16.70	15.76	16.79
9	15.79	16.67	16.60	16.72
14	15.67	16.42	17.15	16.42
19	15.91	16.81	17.21	16.86
24	16.45	16.88	17.13	16.88
29	16.59	17.18	17.08	17.18

Table 5: The result of mean classification error calculated from the different classifiers using feature set obtained from CRA and other traditional methods. Lesser the value means greater robustness.

Conclusion

In this paper, we have proposed a novel correlation and rank aggregation based feature selection method which would help us in identifying most significant features that contribute to accurate prediction of PPE. We have tested our methodology on very high dimensional healthcare data. The results suggest that CRA yields higher prediction accuracy and greater robustness than other traditional feature selection methods that uses distinct statistical properties of data. The experiments suggest that the mean classification error obtained is almost always less than other traditional methods for different size of feature subsets. CRA is helpful not only for high dimensional data applications but also in cases where is it difficult to determine the best statistical property to use for evaluation. The greatest advantage in having a robust technique is that, there will be fewer dilemmas in deciding on the most appropriate classifier from the vast range of choices. CRA had the advantage that, in almost all of the cases features generated from CRA will guarantee to give consistently a higher accuracy. This also shows the significance of the identified variables for detecting PPE. Thus CRA can be considered as a robust and efficient feature selection mechanism especially suited for very high dimensional healthcare data. We conclude that this entire approach has the ability not only to detect the key contributing factors of PPE but also obtain a prediction result with higher accuracy and robustness across different types of classifiers. Our future work includes developing a model that reduces the sparseness of the healthcare data in order to increase prediction accuracy which would help in obtaining a more improved care management for patients with a risk for a Potentially Preventable Event.

Acknowledgement

I would like to thank Tamara Winden and Tammy Lindquist from Allina Health for their advice.

References

[1] Henry J Aaron, Waste, we know you are out there, New England Journal of Medicine 359 (2008), no. 18, 1865–1867.

[2] JC De Borda, Memoire sur les elections au scrutin, 1781, Histoire de lAcademie Royale des Sciences, Paris (1953).

[3] Prasanna Desikan, Nisheeth Srivastava, Tamara Winden, Tammie Lindquist, Heather Britt, and Jaideep Srivastava, Early prediction of potentially preventable events in ambulatory care sensitive admissions from clinical data, Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on, IEEE, 2012, pp. 124–124.

[4] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar, Rank aggregation methods for the web, Proceedings of the 10th international conference on World Wide Web, ACM, 2001, pp. 613–622.

[5] Isabelle Guyon and Andr´e Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003), 1157–1182.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, The weka data mining software: an update, ACM SIGKDD Explorations Newsletter 11 (2009), no. 1, 10–18.

[7] Mark A Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, The University of Waikato, 1999.

[8] Mark A Ilgen, Karen Downing, Kara Zivin, Katherine J Hoggatt, H Myra Kim, Dara Ganoczy, Karen L Austin, John McCarthy, Jignesh M Patel, and Marcia Valenstein, Identifying subgroups of patients with depression who are at high risk for suicide, The Journal of clinical psychiatry 70 (2009), no. 11, 1495.

[9] Kenji Kira and Larry A Rendell, The feature selection problem: Traditional methods and a new algorithm, Proceedings of the National Conference on Artificial Intelligence, John Wiley & Sons Ltd, 1992, pp. 129–129.

[10] Ivan Kojadinovic and Thomas Wottka, Comparision between a filter and a wrapper approach to variable subset selection in regression problems, Proc. European Symposium on Intelligent Techniques (ESIT), Citeseer, 2000.

[11] Mary K Obenshain, Application of data mining techniques to healthcare data, Infection Control and Hospital Epidemiology 25 (2004), no. 8, 690–695.

[12] Sellappan Palaniappan and Rafiah Awang, Intelligent heart disease prediction system using data mining techniques, Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, IEEE, 2008, pp. 108–115.

[13] Industry Report, Smart payment reforms can reduce costs and improve quality:a short primer.

[14] Chandrima Sarkar, Sarah Cooley, and Jaideep Srivastava, Improved feature selection for hematopoietic cell transplantation outcome prediction using rank aggregation, FedCSIS, 2012, pp. 221–226.

[15] Karthik Subbian and Prem Melville, Supervised rank aggregation for predicting influence in networks, arXiv preprint arXiv:1108.4801 (2011).

[16] Frank T¨uttelmann, C Marc Luetjens, and Eberhard Nieschlag, Optimising workflow in andrology: a new electronic patient record and database, Asian journal of andrology 8 (2006), no. 2, 235–241.

[17] Randall Wald, Taghi M Khoshgoftaar, and David Dittman, Mean aggregation versus robust rank aggregation for ensemble gene selection, Machine Learning and Applications (ICMLA), 2012 11th International Conference on, vol. 1, IEEE, 2012, pp. 63–69.

[18] Marc-Oliver Wright, Automated surveillance and infection control: toward a better tomorrow, American Journal of Infection Control 36 (2008), no. 3, S1–S6.

[19] Lei Yu and Huan Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, MACHINE LEARNINGINTERNATIONAL WORKSHOP THEN CONFERENCE-, vol. 20, 2003, p. 856.

* Prof. Jaideep Srivastava, the research supervisor for this study, Consults with Allina Health, the research sponsor. Their relationship has been reviewed and managed by University of Minnesota in accordance with its conflict of interest policies.