DOI: 10.15439/2014F500
ACSIS, Vol. 2

Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets

Eftim Zdravevski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: eftim.zdravevski@finki.ukim.mk

Petre Lameski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: petre.lameski@finki.ukim.mk

Andrea Kulakov
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: andrea.kulakov@finki.ukim.mk

Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: dejan.gjorgjevikj@finki.ukim.mk

***Abstract*—Feature selection is important phase in machine learning and in the case of multi-label classification, it can be considerably challenging. In like manner, finding the best subset of good features is involved and difficult when the dataset has significantly large number of features (more than a thousand). In this paper we address the problem of feature selection for multi-label classification with large number of features. The proposed method is a hybrid of two phases - preliminary feature selection based on the information value and additional correlation-based selection. We show how with the first phase we can do preliminary selection of features from tens of thousands to couple of hundred, and then with the second phase we can make fine-grained feature selection with more sophisticated but computationally intensive methods. Finally, we analyze the ways of allocating the selected features to diverse subsets, which are suitable for training of ensembles of classifiers.**

I. Introduction

MACHINE LEARNING provide means to automatically analyze enormous quantities of data and consequently to: derive various conclusions, make predictions for unseen data, find patterns within the data etc. As learning relies on the available data, its preprocessing is very important to such extent that most of the time of the project might be spent for this phase. During data processing various issues of the data can be addressed: feature modeling and construction [1] [2], outliers removal [3], noise detection and reduction [4], missing values imputation [5] [6], data normalization [7] [8], and data transformation [9] [10].

Many learning algorithms such as neural networks [11], Naive Bayes [12] [13], decision trees [14] notably experience degrading performance when the datasets contain redundant or irrelevant features. This phenomenon is confirmed with theoretical and empirical evidence in plenty of research papers, some of which are [15] [16] and [17]. The problem of feature selection [18] [16] [19] can be defined as the task of selection of subset features that describe the hypothesis at least as well as the original set. The representation of data instances is optimized with feature selection, which in turn can lead to:

- Improving the performance of learning algorithms.
- Reducing the training and execution times of algorithms.
- Improving the memory requirements and allow application of more algorithms.
- Improved robustness to over-fitting.
- Better understanding and visualization of the data.

Different methods for feature selection focus on various aspects of the above goals, or achieve the same goals but in different ways. In [20] are given guidelines for feature selection and are introduced the most widely used methods. It is important to note that finding the most useful and relevant features is not always the same task, as it is shown in [16] and [21].

II. Related work

There are two approaches for feature selection: filter and wrapper approach. The filtering approaches rank the features based on some metric. These methods are generally characterized by simplicity, scalability and solid empirical background. Because they rely on relatively simple metrics, they are memory and computationally efficient and can be applied on datasets with tens or even hundreds of thousands of features. Such application of these methods, as well as their empirical analysis, is further elaborated in [22], [23] and [24]. Filter methods are independent of the machine learning algorithm that is going to be applied later on.

Filter approaches for feature selection can further be categorized into two groups. The first group consists of methods that rank the features based on some measure of their individual predictive power: information value [25] [26] [27], information gain [28] [29], information gain ratio [28] [29], RELIEF [30] [31], entropy [32] etc. In [33] and [34] are described some filtering methods based on posterior probability. The common problem of all methods in this group is that they take into consideration only the individual usefulness of attributes in relation to the target classification and can not discover

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia

978-83-60810-58-3/$25.00 © 2014, IEEE

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

redundancy, multicollinearity or interdependence between the chosen features.

The second type of filter approaches consists of methods which analyze the subset of features based on some metric that describes the performance of the whole subset and not only the individual features [1]. Namely, the correlation-based approaches described in [35] and [36] fall into this type of methods. Important to realize is that they search for subsets of features that have low inter-correlation between them and high correlation to the target classification [37]. Likewise, [38] proposes an approach for detecting stable clusters of features based on principal component analysis.

The wrapper approaches search for subsets of features that are useful for the classification or regression task at hand. They are based on evaluating the performance of different subsets of features using a machine learning algorithm [21] [17] [39]. When applying these methods the individual contribution of features is not being evaluated. In contrast, the contribution of the subset of features is taken into consideration and the whole process is black-box like. In other words, the method does not give exact information why that specific subset of features was selected. In order to apply a particular wrapper method, one has to define: how will be the space of all possible feature subsets traversed; how will be the performance of the learning algorithm evaluated in order to guide the search; and which learning algorithm to be used. If the number of features is small, then all combination of features can be evaluated, but this is rarely the case. The main problem of these methods is their computational complexity. Be that as it may, there are a lot of search techniques that mitigate this problem [19]. On the other hand, the main advantages of these methods is their universality and independence of the domain of the data and task. The research community has proposed various ways of making hybrid methods that combine filter and wrapper and [40] reviews them.

Our research presented in this paper focuses on feature selection in areas of application where datasets have tens or hundreds of thousands of variables. These areas include text processing, gene expression array analysis, and combinatorial chemistry. This paper is organized as follows: Section III describes the problem at hand and section IV gives overview to the proposed solution. In subsections IV-A and IV-B we describe the proposed hybrid approach for feature selection. Subsection IV-C presents various schemes for constructing diverse subsets of features that are suitable for ensembles of classifiers. In Section V we summarize our work.

III. Problem definition

This paper originated from our research during and after the AAIA'14 Data Mining Competition “Key risk factors for Polish State Fire Service”. This competition is organized within the framework of the 9th International Symposium on Advances in Artificial Intelligence and Applications [41], and is an integral part of the 1st Complex Events and Information Modeling workshop devoted to the fire protection engineering. The task is related to the problem of extracting useful knowledge from incident reports obtained from The State Fire Service of Poland. With this in mind, our research goals were mainly guided within the task goals and requirements. Under those circumstances, during the following sections we will occasionally relate to some specifics for this task. Nevertheless, the proposed methods are not specific for this task and they can be applied to a variety of problems.

The organizers obtained a data set containing nearly 260000 reports describing the actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings in years 1992 - 2011. Each report consists of two parts. The first one contains a summary of resources utilized during the action in a form of structured and quantified characteristics. The second part contains a natural language description of the reported events, which is entered by the officer coordinating the action. They have preprocessed a subset of the reports and transformed it into a table in which each of the reports is described by almost 12000 attributes. The training dataset contains about 50000 instances. Additionally, they have distinguished 3 target attributes that correspond to information whether in the described incident there were casualties among firefighters, children or other involved people, respectively. The goal of the competition is participants to come up with solutions which will improve the understanding of the risk factors associated various types of accidents. Given these points, it seems that the problem is actually multi-label classification. As a matter of fact, after careful review of the training data we have observed that some instances (i.e. reports) are indeed classified to the positive classes in more than 1 of the decision attributes. The organizers have modeled the decision attributes in a way that actually transforms the multi-label problem into 3 binary classification problems. Such approach for tackling multi-label problems is, in essence, problem transformation method and is described in [42].

The task in this competition was to identify attributes that can be used to robustly assign the reports to the corresponding decisions labels. In particular, organizers decided to use ensemble of 10 Naive Bayes classifiers for each of the target classifications. Having 3 decision attributes, means that the selected features should be divided into 10 subsets and each subset should be used to train 3 individual Naive Bayes models. Every model assigns scores (i.e. probabilities) to test cases representing if that the case should be classified to the positive decision class or not. In this way, for every decision attribute and every test case there are 10 scores. The ensemble of predictions is constructed by taking the sum of the scores of the individual models.

The metric used to evaluate the performance of the selected attributes was the average AUC of the prediction ensemble for different decision attributes, decreased by a penalty for using a large number of attributes. We assume that the choice of metric is because the data is highly imbalanced and many papers confirm that this metric is best suitable for such cases [43] [44] [45].

Namely, if we denote by: s - submitted solution; |s| a total number of attributes used in the solution (with repetitions); and

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

$AUC_i(s)$ Area Under the ROC Curve (AUC) of a classifier ensemble for the i-th decision attribute, then the quality measure used for the assessment of submissions can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - penalty(s)\right)$$

where the penalty is equal to:

$$penalty(s) = \left(\frac{|s| - 30}{1000}\right)^2$$

and the function F:

$$F(x) = \begin{cases} x, \text{ for } x > 0 \\ 0, \text{ othewise} \end{cases}$$

From all the given task description and stated requirements the following challenges should be acknowledged:

- Evaluation of the usefulness of features in relation to the 3 target classifications.
- Selecting a small subset of features that will be contributing to all 3 target classifications.
- Optimal arrangement of the selected features in N subsets (N=10 in this case) in order to train ensemble of classifiers.

In order to overcome those challenges we propose a hybrid method which is described in the following section.

IV. Proposed method

Selecting the best subsets of features for this dataset is a challenging task because most of the feature selection algorithms cannot be applied due to the large number of features. Additionally some of the methods for feature selection are applicable only on binary classification problems. With this in mind and given that the task at hand has 3 decision attributes, the selection of features that are contributing to the 3 classification tasks at the same time gets even more difficult.

We propose a hybrid approach for feature selection consisting of three phases. The *first phase* performs preliminary feature selection in order to discard the features that are unlikely to contribute to any of the decision classes. The *second phase* applies more sophisticated feature selection algorithms on the dataset that after the first phase has significantly smaller number of features. As a result from the second phase the set of selected features is very concise and all of them contribute to the 3 classification tasks. If the goal was to create 1 model for each of the classification tasks, then we would use the selected features and we use some learning algorithm to build the models. In such case the feature selection would end here. Be that as it may, the contest rules described in III state that the goal is to train an ensemble of Naive Bayes classifiers. Having this in mind, we need a *third phase* that would optimally arrange the chosen features into subsets that will be later used by each individual classifier. We realize that it was not specifically forbidden to use one feature in more than 1 subset. Although this may be allowed, we believe that such approach is problem-specific and would require a significant effort for fine tuning, to the extent that the scientific contribution of the approach would diminish. For this reason we have decided to use diverse subsets of features for each individual classifier. In other words, each selected feature belongs to only 1 subset. The following subsections describe each of the phases in our approach.

A. Preliminary feature selection

The large number of feature in the original dataset presents a difficult task for most methods. The reason for this is because of the memory and/or computational complexity it imposes. The goal of this task is to overcome that problem by reducing the features to a significantly smaller number using some simple algorithm. Being able to do this clears the way for more sophisticated feature selection methods. As it was explained in section I, the prime candidates for a fast preliminary (i.e. coarse-grained) feature selection are the filter methods that assess that individual contribution of features. The following metrics can be used for feature selection are less demanding in terms of memory and computational power: information gain, information gain ratio [28] [29] and information value [25] [26] [27]. In spite of the slight differences between them in terms of computational complexity time, all of them can be computed in linear time ($O(mn)$) with 1 pass of the training dataset. We were not able to obtain results with the RELIEF method [30] [31] in reasonable time due to its higher complexity - $O(mnp)$. Here where m is number of training instances, n is the number of attributes and p is the number of randomly selected instances used for the RELIEF algorithm. We acknowledge that with proper tuning of the p parameter we might have been able to obtain results with it too, but since this phase performs only preliminary selection we believe that this is not worth the effort.

We have decided to use the information value for estimation of the predictive power of each of the features in relation to each of the decision attributes. It is widely adopted in industry especially for credit scoring problems [25] [26] [27]. The reason for this is because there are some widely adopted rules of thumb in terms that give simple guidelines of whether the feature is strong or weak predictor based on the information value. However, note that weak features may provide value in combination with others; or have individual values that could provide predictive power as dummy variables. As it has been suggested in [46], the following guidelines for evaluating the strength a predictor based on the information value can be used:

- Less than 0.02: unpredictive
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- Greater than 0.3: strong

Although they are firmly grounded in good practice, how these guidelines be related to other metrics is discussed in [47]. At the same time, there are some drawbacks of this metric related to the some border cases that prevent using its original

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

definition (2). In [10] are proposed some enhancements of the weight of evidence (WoE) parameter, which in turn overcome the computational obstacles for the information value. With (1) is defined WoE, and it is further used for calculation of the information value (2). Here N_i^j and P_i^j represents the number of negative and positive instances labeled with the i-th value of the j-th feature, respectively. Also SN and SP denote the total number of negative instances and the total number of positive instances in the training dataset, respectively.

$$WoE_i^j = ln\left(\frac{N_i^j}{P_i^j}\right) - ln\left(\frac{SN}{SP}\right) \tag{1}$$

$$IV^j = \sum_{i=1}^{n}\left[\left(\frac{N_i^j}{SN} - ln\frac{P_i^j}{SP}\right) \times WoE_i^j\right] \tag{2}$$

From (2) it is obvious that the information value is applicable only to binary classification problems. When having multi-label classification with k possible positive labels, one needs to compute k information values for each of the features. For this specific task, having 3 decision attributes and almost 12000 features, means that we had to compute nearly 36000 information values. The computation for all of them takes less than 15 minutes on a regular laptop. As it turns out, some of the features are strong or medium predictors for one decision attribute, but are very weak predictors for the other one or two target attributes. The next challenge was how to aggregate the 3 information values of each feature into 1 value, so we can use it for feature selection. The following subsections describe the results of each aggregation type.

1) Average information value: When having multi-label classification with k positive classes we can average the k information values of each feature to get estimate of the its information value in relation to all positive classes. For this case in particular, we have tried averaging the 3 information values of each feature in order to use it for feature selection. We have examined various subsets containing 50 to 120 of the best features based on their average information value. By training ensembles of Naive Bayes classifiers as described in III we have obtained AUC performance on the leader-board dataset varying from 0.886 to 0.9, based on the number of features and the scheme of arrangement of them into diverse subsets. As a reference, the best results of the same test dataset were up to 0.94. The experiments showed that the performance of the ensembles build on the these selected features were fairly stable. However, they were worse than what we hoped to be achieved with more sophisticated methods. Nevertheless, it was notable that the average information value can be safely used for preliminary feature selection.

2) Maximum information value: The next obvious idea for aggregating the individual information values of a feature is to calculate their maximum. When we applied this logic on the current dataset and we selected the best different subsets containing 50 to 120 features based on their maximum information value, the performance of the ensembles was worse than with the approach in IV-A2. In fact, the AUC performance on the leader-board dataset was less than 0.8, regardless of the arrangement of the features in subsets. By looking into the selected features and their maximum information values we can explain this phenomenon. As it can be observed on Fig. 1, some features might have high information value for one of the decision attributes, but low for the other decision attributes. This in turn, translates to high maximum but low average information value.

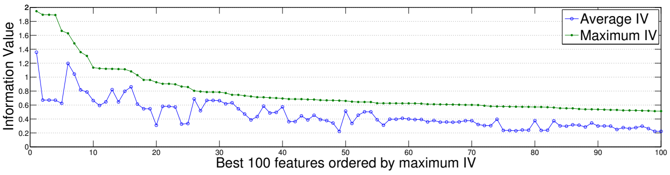

Fig. 1. Maximum vs. average information value of best 100 features

To conclude, this aggregation might be bad-performing for final feature selection, but one should not rush to avoid it. In fact, this aggregation may identify features for training models for each of the decision attributes separately and may also identify features that in combination with others might be very useful. The maximum information value is very useful for discarding features because it guarantees that the discarded ones are bad or week predictors for all decision attributes (positive labels).

3) Weighted average information value: Another approach for aggregating the individual information values of a feature is to calculate their weighted arithmetic average. The weight can be calculated based on the number of positive instances in the training dataset for each decision attribute (i.e. label). This idea has been applied for weighting averages of other statistics [48] [49]. However, for final feature selection wighted average type of aggregation does not seem to be suited, mostly because the most common positive label is always preferred. On the leader-board dataset the AUC of different subsets varying from 50 to 120 features was about 0.86. Maybe with a different weighting scheme, the performance would be improved. However, as this phase should not select final subset of features, we did not investigate other weighting schemes.

4) Dependent features: It has been extensively proved that features with high correlation have negative impact on performance for many machine-learning algorithms, among which is the Naive Bayes classifier. Some such papers are mentioned in section I. In order to address this issue, we have calculated the correlation coefficients and p values [50] between all features and used this information to find dependent features. By discarding a feature if we have already selected a dependent feature with higher aggregate information value, we were able to slightly improve the performance of the maximum and weighted average aggregations. In those cases, the performance was similar to the average aggregation. Despite that, the obtained results were not satisfactory for final feature selection.

5) Coarse-grained selection of features: We have applied the three aggregation methods for the information values and

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

then we selected sets of the best N features (where N is 400, 500, 600, 800, 1000, and 1500). By analyzing the sets obtained for different values for N we have concluded that for the same value of N but different aggregation methods most of the features (i.e. 70-90 %) are overlapping. That indicated the aggregation type will not have significant impact on the coarse-grained selection of features, except for their ranking. In order to decide how many features to select during this phase we have analyzed the max-aggregated information values and have noticed that after the best 500 features the maximum information value drops bellow 0.1, meaning that the discarded features are week predictors for all decision labels. The maximum information value is very useful for discarding features because it guarantees that the discarded ones are bad or week predictors for all decision attributes (positive labels).

Finally, as a result from this phase we have selected the best 500 features based on the maximum information value and we continued with the next phase to apply more sophisticated feature selection algorithms on the significantly simplified training dataset.

B. Fine-grained feature selection

After phase 1, the training dataset is significantly simplified. More specifically for the current dataset, we have 50000 instances and 500 features. It is significantly reduced than the original, so more intelligent feature selections algorithms can now be applied.

The number of features in a dataset should indicate whether it is possible to use wrapper methods for fine-grained feature selection. The experiments performed in [51] show that wrapper methods can be applied to relatively smaller datasets (containing less than 200 features and few thousand instances). In spite of the continued improvement in processing power during the recent years, trying out many combination of features especially with more complex learning algorithms is a very hard task. In this case, the reduced size of dataset is still quite large in order to be able to apply a wrapper method for feature selection on it. This fact limited the use of wrapper method to very simple internal learning algorithms. Again, having too many combinations of feature subsets still makes wrapper methods not adequate for this task.

One of the best performing methods for evaluation of subsets of features is the correlation-based feature subset selection [35]. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred.

If we were to apply this method on the 3 decision attributes separately, we would still need to aggregate the 3 selected subsets of features. As it was shown in subsection IV-A, this task can be very involved. Instead of doing that, we have decided to transform the problem space. The original task is multi-label classification which was transformed by the organizers to 3 separate binary classification problems. In cases like this, we propose to merge the separate binary classification problems into 1 multi-class problem. To summarize, starting with a multi-label problem transformed as several separate binary problems we merge it to a multi-class problem. By doing this we can apply feature selection methods that select the best features in relation to all positive classes. More particularly, with the proposed transformation we obtained 8-class classification problem by using the following Eq. (3) to map each instance to a new artificial class. Here, AL_i denotes the artificial label of the multi-class problem instance i, where as L_i^1,L_i^2 and L_i^3 are the classes in the binary classifications of the same instance.

$$AL_i = 1 \times L_i^1 + 2 \times L_i^2 + 4 \times L_i^3 \tag{3}$$

In general, multi-label classification tasks where the number of positive labels is N, can be transformed to N binary classification problems [42]. Let the label of the i-th instance in the $j-th$ binary problem is L_i^j, where for L_i^j is 0 for negative instances and 1 for positive instances. With this transformation to multi-class problem the same instance will be labeled with AL_i as defined in Eq. (4):

$$AL_i = \sum_{j=1}^{N} 2^{j-1} \times L_i^j \tag{4}$$

After performing this transformation, the correlation-based feature subset selection can be applied. Depending on how many features are in the training dataset and how they are chosen (i.e. which aggregation was used), this method selects from 40 to 70 features. Considering the obtained attributes we have observed that one particular subset of 53 features was very common, henceforth the next phase was performed using that subset (shown on Table I).

C. Allocation of features into diverse subsets

After end of phase 2 we have a very concise dataset which, in this case, is described with 53 features. The correlation-based method for feature selection [35] does not rank the features, but we can rank them computed based on the information value calculated during phase 1. They can be ranked based on their maximum or average information value.

The goal of this phase is to optimally allocate the selected features into diverse subsets. For this task the number of subsets is set at 10, but in general, one can try various number of individual classifiers for the ensemble. Each subset should contain approximately equal number of features. The following subsections describe the schemes for allocation of features into subsets. Before we continue, let us define an *iteration* as allocating 1 feature to each subset (e.g. in this case choosing 10 features, 1 for each subset). The different schemes explained bellow, have different logic of choosing the next feature to allocate to a subset. If we consider the subsets as items that are ordered, we can decide which of them will get processed first. By being processed we mean allocating a feature to it. Likewise, the features are ordered by their maximum information value.

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

1) FIFO scheme: The First-In-First-Out (i.e. First-Come-First-Served) term has been widely used in data structures literature and queue theory. During 1 iteration the FIFO scheme would allocate the next best feature to the next subset. The following iteration will allocate features starting from the first subset and so on until there are no more features. Obviously, this scheme mostly favors the first subset and least favors the last. Using the leader-board dataset, we have obtained AUC of 0.9292. To summarize, this approach uses the maximum information values for ranking the features. When we used the average information value for ranking, the performance was slightly worse. The simple explanation for this is because the average information value is more consistent than the maximum, hence the FIFO scheme favors the first subsets more.

2) FIFO-independence scheme: In order to improve the FIFO scheme, we can dependent features in order to optimally allocate the features into subsets. The idea is to have independent features within 1 subset. The algorithm used during phase 2 is correlation based which ensures that the selected features have very low inter-correlation among them. However, if we use a more strict test for independence (p value = 0.01), then we can still find some pairs of dependent features. This improved FIFO scheme selects the next best feature that is independent to all features that are already in the subset. As it turns out, when using the leader-board dataset, this scheme slightly improved the performance to AUC of 0.9293.

3) Interchanging FIFO-FILO scheme: With this scheme in each iteration we change the logic from FIFO to FILO and vice versa. So, when assigning the first attribute to the first subset we choose the best feature, then for the second subset we choose the next best feature and so on, until the last subset has 1 feature. Then when assigning a second feature to all subsets, we start with the last subset and assign the best available feature to it. In like manner, we continue assigning the next best feature until the first subset has 2 features. In the next iteration the first subset will have priority, and so on until we run out of features. With this scheme the AUC ROC performance on the leader-board dataset was 0.9298, which was an additional improvement.

4) Monte Carlo scheme: This scheme randomly scatters the features to subsets. It is the simplest scheme and produced results ranging from 0.926 to 0.9321. We have analyzed the final distributions to subsets that produced better and the ones that produced worse results. When looking at the information values for all three target attributes of the features in each subset it was notable that the better performing arrangements of subsets had features that are medium or strong predictors in relation to 1 or 2 target attributes and weak predictors for the other target attribute.

We have concluded that this scheme might produce very good results, but in order to be consistent it needs to be improved. One way of doing this is to use this scheme as a starting point and later to make rearrangements by swapping some features between the subsets. Choosing which features to swap is based on the following logic:

- We first find a bad performing subset of features and determine which target attribute has least weak features in relation to it.
- Then find a subset where a lot of features are medium or strong predictors for the same target attribute.
- Swap the features from the 2 subsets.
- Repeat the process until no swaps can be made.

This algorithm generally helps both subsets. The first subset will get a stronger feature for the class that has bad performance. The second subset is also improved because the possibility of over-fitting because of too many strong predictors for it for particular target attribute is reduced. Using this technique we have finally arrived at the feature arrangement shown on Table I on the following page.

V. Conclusion

In this paper we have proposed a three-phase hybrid feature selection method that is able to extract features from datasets with thousands of features. This method is especially useful for datasets that originate from text processing areas of application. Additionally we have analyzed the different ways to aggregate information values of one feature in the case of multi-label classification. As a consequence we have pointed out the advantages and shortcomings of the aggregation types. Also we have proposed and analyzed different schemes of allocation of the selected features to diverse subsets that are suitable for training ensembles of classifiers. Equally important was the proposed method of transforming multi-label classification problems into multi-class in order to be able to apply some feature selection algorithms. We have tested the proposed methods on the AAIA'14 data mining competition dataset [41] and our solution has been recognized as one of the top 5.

References

[1] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection a Data Mining Perspective*. Boston, MA: Springer US, 1998. ISBN 9781461557258 1461557259. [Online]. Available: http://dx.doi.org/10.1007/978-1-4615-5725-8

[2] C. J. Matheus and L. A. Rendell, "Constructive induction on decision trees," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'89. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 645–650. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623755.1623857

[3] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should always check for them)," *Practical assessment, research & evaluation*, vol. 9, no. 6, pp. 1–12, 2004.

[4] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber, "On noise reduction methods for chaotic data," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 3, no. 2, 1993.

[5] R. J. A. Little, *Statistical analysis with missing data*, 2nd ed., ser. Wiley series in probability and statistics. Hoboken, N.J: Wiley, 2002. ISBN 0471183865

[6] P. Royston, "Multiple imputation of missing values," *Stata Journal*, vol. 4, pp. 227–241, 2004.

[7] A. A. Hancock, E. N. Bush, D. Stanisic, J. J. Kyncl, and C. Lin, "Data normalization before statistical analysis: keeping the horse before the cart," *Trends in Pharmacological Sciences*, vol. 9, no. 1, pp. 29 – 32, 1988. doi: http://dx.doi.org/10.1016/0165-6147(88)90239-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165614788902398

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I
TABLE OF SELECTED FEATURES AND THEIR INFORMATION VALUES

Index	Subset	IV 1	IV 2	IV 3	Average IV	Max IV
3001	1	0.0700	0.3834	0.2928	0.2487	0.3834
258	1	0.0041	0.0612	0.2769	0.1141	0.2769
11463	1	0.2804	0.4048	0.4848	0.3900	0.4848
5041	1	0.0040	0.1496	0.1617	0.1051	0.1617
1880	1	0.4021	0.0992	0.1233	0.2082	0.4021
5996	1	0.0504	0.1541	0.2184	0.1410	0.2184
4415	2	0.1306	0.1852	0.0843	0.1334	0.1852
4698	2	0.0105	0.3537	0.3403	0.2348	0.3537
3027	2	0.1437	0.1611	0.1555	0.1534	0.1611
10428	2	0.0835	0.2628	0.4815	0.2759	0.4815
1772	2	0.1626	0.7043	0.5911	0.4860	0.7043
3955	2	0.0086	0.2998	0.2601	0.1895	0.2998
1509	3	0.0005	1.3586	1.0919	0.8170	1.3586
143	3	0.5076	1.4528	1.6283	1.1962	1.6283
8114	3	0.0443	0.6701	0.4605	0.3916	0.6701
1538	3	0.0236	0.3462	0.3900	0.2533	0.3900
7999	3	0.0372	0.1050	0.2726	0.1383	0.2726
7425	3	0.6674	0.2486	0.2163	0.3774	0.6674
5270	4	1.6636	0.0854	0.1233	0.6241	1.6636
460	4	0.0339	0.3748	0.4775	0.2954	0.4775
11701	4	0.3482	1.7748	1.9455	1.3562	1.9455
142	4	0.0051	1.8940	0.1132	0.6708	1.8940
11165	4	0.6436	0.3584	0.3611	0.4544	0.6436
139	5	0.0024	0.9596	0.6780	0.5467	0.9596
7055	5	0.4033	0.0169	0.0846	0.1683	0.4033
1335	5	0.0127	0.1994	0.0969	0.1030	0.1994
11825	5	0.3693	0.6592	0.5104	0.5130	0.6592
8635	5	0.3322	1.1180	1.0114	0.8205	1.1180
6660	6	0.1667	0.3141	0.3896	0.2902	0.3896
11459	6	0.0142	0.1378	0.1587	0.1036	0.1587
10771	6	0.3630	0.1023	0.1219	0.1957	0.3630
6779	6	0.0035	1.1367	0.8519	0.6640	1.1367
9657	6	0.2205	0.2340	0.1535	0.2027	0.2340
5306	7	0.4444	0.0590	0.0870	0.1968	0.4444
11100	7	0.0297	0.6239	0.5207	0.3914	0.6239
10638	7	0.5431	0.2341	0.1554	0.3109	0.5431
1244	7	0.1110	1.1247	0.5446	0.5934	1.1247
7187	7	0.0051	1.8906	0.1126	0.6694	1.8906
5909	8	0.0015	1.3041	1.0542	0.7866	1.3041
4210	8	0.0002	0.1624	0.1548	0.1058	0.1624
7007	8	0.1164	0.1948	0.1086	0.1400	0.1948
1767	8	0.0197	0.1962	0.1891	0.1350	0.1962
1152	8	0.2480	0.2208	0.1964	0.2217	0.2480
5925	9	0.0061	0.5386	0.4829	0.3425	0.5386
8039	9	0.5741	0.1034	0.0517	0.2431	0.5741
2182	9	1.0820	0.6352	0.8652	0.8608	1.0820
8073	9	0.5194	0.1280	0.2433	0.2969	0.5194
3257	9	0.6396	0.7863	0.5737	0.6665	0.7863
6162	10	0.0629	0.5856	0.6844	0.4443	0.6844
8487	10	0.1549	0.4232	0.3095	0.2959	0.4232
8914	10	0.0292	0.2425	0.0870	0.1196	0.2425
10968	10	0.1452	0.3279	0.3999	0.2910	0.3999
1038	10	0.2116	0.3394	0.3006	0.2839	0.3394

[8] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *Nuclear Science, IEEE Transactions on*, vol. 44, no. 3, pp. 1464–1468, Jun 1997. doi: 10.1109/23.589532

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*. Menlo Park, Calif.: AAAI Press : MIT Press, 1996. ISBN 0262560976 9780262560979

[10] E. Zdravevski, P. Lameski, and A. Kulakov, "Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, July 2011. doi: 10.1109/IJCNN.2011.6033219. ISSN 2161-4393 pp. 181–188.

[11] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, 3 1997. ISBN 9780070428072. [Online]. Available: http://amazon.com/o/ASIN/0070428077/

[12] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN 1-55860-612-2 pp. 258–267. [Online]. Available: http://dl.acm.org/citation.cfm?id=645528.657649

[13] R. O. Duda, *Pattern classification*, 2nd ed. New York: Wiley, 2001. ISBN 0471056693

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0

[15] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'91. AAAI Press, 1991. ISBN 0-262-51059-6 pp. 547–552. [Online]. Available: http://dl.acm.org/citation.cfm?id=1865756.1865761

[16] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1âĂŞ2, pp. 245 – 271, 1997. doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5 Relevance. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370297000635

[17] P. Langley, *Elements of machine learning*. San Francisco, Calif: Morgan Kaufmann, 1996. ISBN 1558603018

[18] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proceedings of the Eleventh International Conference*. Morgan Kaufmann, 1994, pp. 121–129.

[19] B. Raman and T. R. Ioerger, "Instance based filter for feature selection," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 1–23, 2002.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[21] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997. doi: 10.1016/S0004-3702(97)00043-X. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[22] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional word clusters vs. words for text categorization," *J. Mach. Learn. Res.*, vol. 3, pp. 1183–1208, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944969

[23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944974

[24] L. Hermes and J. Buhmann, "Feature selection for support vector machines," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, 2000. doi: 10.1109/ICPR.2000.906174. ISSN 1051-4651 pp. 712–715 vol.2.

[25] R. Anderson, *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press, 2007. ISBN 9780199226405

[26] S. Finlay, *Credit scoring, response modeling, and insurance rating: a practical guide to forecasting consumer behavior*, 2nd ed. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan, 2012. ISBN 9780230347762

[27] N. E. Mays, Lynas, *Credit scoring for risk managers: the handbook for lenders*. S.l.: CreateSpace], 2010. ISBN 9781450578967 1450578969

[28] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 155–165, Jan. 2006. doi: 10.1016/j.ipm.2004.08.006. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2004.08.006

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[30] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. ISBN 1-5586-247-X pp. 249–256. [Online]. Available: http://dl.acm.org/citation.cfm?id=141975.142034

[31] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Machine Learning: ECML-94*, ser. Lecture Notes in Computer Science, F. Bergadano and L. De Raedt, Eds. Springer Berlin Heidelberg, 1994, vol. 784, pp. 171–182. ISBN 978-3-540-57868-0. [Online]. Available: http://dx.doi.org/10.1007/3-540-57868-4_57

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.

[32] T. Jebara and T. Jaakkola, "Feature selection and dualities in maximum entropy discrimination," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. ISBN 1-55860-709-9 pp. 291–300. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073946.2073981

[33] A. Vehtari and J. Lampinen, "Bayesian input variable selection using posterior probabilities and expected utilities," *Report B31*, 2002.

[34] A. Y. Ng and M. I. Jordan, "Convergence rates of the voting gibbs classifier, with application to bayesian feature selection," in *In 18th International Conference on Machine Learning*. Morgan Kaufmann, 2001.

[35] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[36] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[37] M. Dash, H. Liu, and H. Motoda, "Consistency based feature selection," in *Knowledge Discovery and Data Mining. Current Issues and New Applications*, ser. Lecture Notes in Computer Science, T. Terano, H. Liu, and A. Chen, Eds. Springer Berlin Heidelberg, 2000, vol. 1805, pp. 98–109. ISBN 978-3-540-67382-8. [Online]. Available: http://dx.doi.org/10.1007/3-540-45571-X_12

[38] A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," in *Functional Genomics*, ser. Methods in Molecular Biology, M. Brownstein and A. Khodursky, Eds. Humana Press, 2003, vol. 224, pp. 159–182. ISBN 978-1-58829-291-9. [Online]. Available: http://dx.doi.org/10.1385/1-59259-364-X%3A159

[39] P. Yang, W. Liu, B. Zhou, S. Chawla, and A. Zomaya, "Ensemble-based wrapper methods for feature selection and class imbalance learning," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, J. Pei, V. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Springer Berlin Heidelberg, 2013, vol. 7818, pp. 544–555. ISBN 978-3-642-37452-4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37453-1_45

[40] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. ISBN 1-55860-778-1 pp. 74–81. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.658297

[41] "Aaia'14 data mining competition, howpublished = https://fedcsis.org/2014/dm_competition, note = Accessed: 2014-05-30."

[42] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084 – 3104, 2012. doi: http://dx.doi.org/10.1016/j.patcog.2012.03.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312001203

[43] A. P. Bradley, "The use of the area under the {ROC} curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997. doi: http://dx.doi.org/10.1016/S0031-3203(96)00142-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320396001422

[44] C. X. Ling, J. Huang, and H. Zhang, "Auc: A statistically consistent and more discriminating measure than accuracy," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, ser. IJCAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 519–524. [Online]. Available: http://dl.acm.org/citation.cfm?id=1630659.1630736

[45] J. Huang and C. Ling, "Using auc and accuracy in evaluating learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 299–310, March 2005. doi: 10.1109/TKDE.2005.50

[46] N. Siddiqi, *Credit risk scorecards: developing and implementing intelligent credit scoring*. Hoboken, N.J: Wiley, 2006. ISBN 9780471754510

[47] L. Bruce and D. Brotherton, "Information value statistic," in *Midwest SAS User Group 2013 Conference Proceedings*. Marketing Associates, LLC, 2013, pp. 1–18.

[48] A. Rrnyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[49] P. J. Fleming and J. J. Wallace, "How not to lie with statistics: The correct way to summarize benchmark results," *Commun. ACM*, vol. 29, no. 3, pp. 218–221, Mar. 1986. doi: 10.1145/5666.5673. [Online]. Available: http://doi.acm.org/10.1145/5666.5673

[50] M. J. Schervish, "P values: what they are and what they are not," *The American Statistician*, vol. 50, no. 3, pp. 203–206, 1996.

[51] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering," in *Advances in Intelligent Data Analysis VI*, ser. Lecture Notes in Computer Science, A. Famili, J. Kok, J. Pena, A. Siebes, and A. Feelders, Eds. Springer Berlin Heidelberg, 2005, vol. 3646, pp. 440–451. ISBN 978-3-540-28795-7. [Online]. Available: http://dx.doi.org/10.1007/11552253_40

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on January 11,2025 at 19:17:34 UTC from IEEE Xplore. Restrictions apply.