ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Feature Selection and Ensemble Clustering Mechanism for High Dimensional Imbalanced Class Data Using Harmony Search Technique.

Avhad Ankita V., Dr. Kshirsagar D. B.

Department of Computer Engineering, SRES Sanjivani College of Engineering, Kopargaon , India.

H. O. D, Department of Computer Engineering, SRES Sanjivani College of Engineering, Kopargaon , India

ABSTRACT: The recent increase of data poses a severe challenge in data extracting. High dimensional data can contain high degree of irrelevant and redundant information**.** In real-world applications, the cost of an incorrect classification of data for minority classes can be very high. This is a challenging problem, especially when the data are too high at the highest level of the dimension due to the increase in the feet and the interpretation of the lower model. Selection of functions recently identified by the characteristics that best fit a minority class. This has become a popular way to solve the problem. This document introduces a new method of selecting activities called SYMON that uses symmetric uncertainty. It makes the search for harmony. Unlike current methods, SYMON uses symmetric uncertainty to evaluate characteristics in relation to their dependence on the class label. Less often it helps to identify powerful features in class label recovery. SYMONuses harmony search to prepare the feature selection phase as an optimization problem to choose the best possible combination of functionality. The proposed algorithm is able to handle situations in which a set of attributes has the same weight, which includes two vector tuning operations involved in the harmony search process. This document has expanded the ideas to reduce the piece of the classifier, changing the prediction of the dress in the training samples and for problems of selection of the convenience for the treatment of the classifier as features.

KEYWORDS: Feature Selection, Ranked Features, Ensemble Clustering, Symmetric Uncertainty, Harmony Search, Wrapper Method

I. INTRODUCTION

In recent years, social media services are used very widely emerging that allow people to communicate and express themselves conveniently and easily. The huge use of social media generates massive and high dimensional data. This poses new challenges to the task of data mining such as classification and clustering processes. One approach for handling such large scale and high dimensional data is Feature Selection. Feature selection methods have been used form long years in the field of statistics and pattern recognition with the wide spread use of machine learning techniques [1]. Feature selection methods are needed when there are too much data that can be processed efficiently by machine learning algorithms, or when some features are costly to acquire and hence the minimum number of features are preferred [4]. The presence of imbalanced data is a problem for classification algorithms. An imbalanced data set is one where at least one class is under-represented compared to the others. Such data creates many challenges to the process of knowledge discovery and has many implications in real-world applications. Addressing these issues brings about many good solutions, such as MIROS1 that is used to detect the possibility of oil spilling, or to detect malicious activities of users in the context of network intrusion as seen in the AIDE environment. This paper investigate the imbalanced class problem further by considering cases where the data set is also high in dimensionality ,thus making the problem more pronounced as the efficacy of learning algorithms is further reduced. Different approaches have been proposed to address the imbalanced learning problem, including resampling one-class learning cost-sensitive learning and feature selection (Yin et al., 2013; Maldonado et al., 2014; Alibeigi et al., 2012). In resampling, the two most

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

common techniques used for the imbalanced data problem are (i) random oversampling and (ii) random under sampling (Van Hulse et al., 2009; Maldonado et al., 2014). In the former, random duplicates of instances from the minority class are added to the original data set, leading to longer classifier training time. With the latter, the instances from the majority classes are randomly discarded, thus the information loss usually leads to sub-optimal learning outcomes. One example of random oversampling is Synthetic Minority Oversampling TEchnique (SMOTE) proposed by Chawla et al. (2004) and Deepa and Punithavalli (2011). The algorithm generates artificial examples for the minority class by interpolating the current minority instances and has been shown Anil Kumar and Ravi (2008) to improve classification performance over imbalanced data.

II. RELATED WORK

The paper [1] describes Density Based Feature Selection (DBFS) is that features' distributions over classes can bring significant benefits to feature selection algorithms.DBFS takes into account features' corresponding distributions over all classes along with their correlations. Advantages are: DBFS has the highest performance across AUC and F1 evaluation statistics over all data sets.To classify instances more accurately. Disadvantages are: It does nothandle multi-class problems. The paper [2] elaborates a new One-Versus-Each (OVE) framework; a rule has to be relevant for one class and irrelevant for everyother class taken separately.The approach, called fitcare, is experimentally validated on various benchmark data sets and our theoretical findings are confirmed. Advantages are: Fitcare's emphasis on correctly predicting minority classes, it also competes with the state-of-the-art associative classifiers in terms of global performance.Fitcare provides the best per-class accuracy.To automatically tune the parameters so that the confusion between any pair of classes is minimized. Disadvantages are: The cost-sensitive classification is required.

The paper [3] presents novel document clustering algorithms based on the Harmony Search (HS) optimization method.Harmony clustering is constituted with the K-means algorithm in three ways to achieve better clustering by combining the explorative power of HS with the refining power of the K-means. Advantages are: Harmony search to get close to optimal solution.K-means algorithm fine tunes that.Higher cluster quality.Disadvantages are: Increases time complexity.A new variation of ant colony optimization (ACO) [4] that utilizes an intelligent method for selection of edges and updating the pheromone of solutions to better guide the search process. The proposed algorithm is referred to as RACO.RACO is applied to the task of feature selection (RACOFS) to show the effectiveness of the algorithm in its application. Advantages are: To increase the exploration and exploitation abilities of the ants and correspondingly prevent the algorithm converging prematurely.RACOFS, showed significant superiority in both the KS and CA measures.

Proposes an efficient Web page recommender by exploiting session data of users.A novel clustering algorithm [5] to partition the binary session data into a static number of clusters and utilize the partitioned sessions to make recommendations. Advantages are: Binary clustering algorithm is scalable.Better clustering quality by combining explorative power with fine-tuning power of the k-means algorithm.Accuracy is high.Disadvantages are: Does not handle cold-start and dynamic pages.Does not the compute the similarity of sessions. The paper [6] proposes an automatic music genre-classification system based on a local feature selection strategy by using a self-adaptive harmony search (SAHS) algorithm.A feature-selection model using the SAHS algorithm is then employed for each pair of genres, thereby deriving the corresponding local feature set. Advantages are: Proposed method is more effective than other relevant methods.The local feature selection still performs better than the global feature selection.SAHS algorithm could be also applied to various applications such as movie genre classification and painting genre classification.

Proposes a two-step approach in paper [7] that identifies a set of candidate features based on the data characteristics and then selects a subset of them using correlation and instance-based feature selection methods.Apply three advanced machine learning feature selection algorithms - Mutual Information (MI), RReliefF (RF) and Correlation-Based Selection (CFS) – to the task of load forecasting. Advantages are: Ability to model non-linear relationships between the predictor variables and the output variable.Ability to learn from examples and extract patterns, instead of making assumptions about the process that generated the data and fitting this model.Noise tolerance. Disadvantages are: It does not select the best feature set.A backward elimination approach [8] for feature ranking and embedded classification uses Support Vector Machines, which has been adapted to select those attributes

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

that are relevant to discriminate between classes under imbalanced data conditions.The order of the algorithm for BFE-SVM with balanced loss is exactly the same as that of RFE-SVM, and it may lead to better results by defining an adequate loss function for classification performance with imbalanced data sets. Advantages are: The proposed approaches outperform other feature ranking techniques in terms of predictive performance for different SVM-based feature selection techniques, achieving particularly good results on highly imbalanced data sets, based on their ability to identify irrelevant variables using the classifier and minimizing the number of errors in the minority class, which is assumed to have a higher cost.The proposed methods allow for the explicit incorporation of misclassification costs in the assessment of each attribute's contribution, leading to a feature selection process especially designed for a particular application.Our strategies are very flexible and allow the use of different kernel functions for nonlinear feature selection and classification using SVM. Furthermore, they can also be generalized to various classification methods, other than SVM. Disadvantages are: Need for an intelligent oversampling in extreme cases of class imbalance and overlap,in which no adequate classifier can be found, since embedded and wrapper feature selection strongly depend on the classification method.

The paper [9] proposes a hybrid search method based on both harmony search algorithm (HSA) and stochastic local search (SLS) for feature selection in data classification.A novel probabilistic selection strategy is used in HSA–SLS to select the appropriate solutions to undergo stochastic local refinement, keeping a good compromise between exploration and exploitation. Advantages are:Simplicity, flexibility and robustness.The HSA–SLS with the probabilistic selection strategy is effective for feature selection and classification.Disadvantages are: Real-world problems do not use in order to verify and extend the proposed method. The paper [10] proposes a new document frequency and term frequency combined feature selection method (DTFS).TFSM combined an optimal DFFS method (ODFFS) and a proposed novel TFFS method (NTFFS) to select the most discriminative features.The harmony search method by introducing a factor, namely best harmony considering rate (BHCR), to search the optimal thresholds.Advantages are: HS is free from divergence.HS does not require initial value settings of the decision variables, thus it may escape the local optima.HS generates a new vector, after considering all of the existing vectors, whereas the other algorithms. Disadvantages are: E-mail classification technologies mainly focus on the text information in feature selection process does not apply on non-text information.

A novel feature selection algorithm, which is governed by biological knowledge, is developed. Gene expressiondata being high dimensional and redundant, dimensionalityreduction is of prime concern. We employ the algorithm clustering large applications based on Randomized search (CLARANS)for attribute clustering and dimensionality reduction based on geneontology (GO) study. Feature selection with unsupervised learningis a difficult problem, with neither class labels present nor anyguidance available to the search. Determination of the optimalnumber of clusters is another major issue, and has an impact onthe resulting output.

This results in dimensionality reduction, withparticular emphasis on high-dimensional gene expression data,thereby helping one to focus the search for meaningful partitionswithin a reduced attribute space. While most clusteringalgorithms require user-specified input parameters, it is oftendifficult for biologists to manually determine suitable values forthese. The use of clustering validity indices for an automateddetermination of optimal clustering has been reported in the literature. In this paper, we incorporate biologicalknowledge, in terms of GO, along with the *DB* index, to automaticallyextract the biologically relevant cluster prototypes.

III. SYSTEM DESIGN

Feature selection takes a different view by shifting the focus to the features (i.e. dimensions) rather than the training examples. The key idea is to find a subset of features that optimize the contrast between classes in the data.

The wrapper and embedded approaches were proposed to produce a more targeted feature subset. These approaches can be based on the ranking of features, where the criteria are often a loss function, e.g. the contribution of a feature to the classification rate, or the discriminative power of features. Selecting features based on a loss function does not always yield the best learning outcome for the classifier. Rather, ranking features with respect to their dependency towards a class label and using that information to select the feature subset would give better performance, especially in predicting the minority class. Proposes a SYMON algorithm for solving problem:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Propose novel approach to ensemble clustering for better result acquisition. For that we aims to generating a set of diverse ensemble members, while the objective of the second stage is to select a suitable consensus function to summarize the ensemble members and search for an optimal unified clustering solution.
We propose an incremental ensemble framework for semi supervised clustering in high dimensional feature spaces. Second, a local cost function and a global cost function are proposed to incrementally select the ensemble members. Third, the newly designed similarity function is adopted to measure the extent to which two sets of attributes are similar in the subspaces. Fourth, we use non-parametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets.

Advantages of Proposed System:

Advantages:

1. The incremental ensemble member selection process is a general technique which can be used in different semi-supervised clustering ensemble approaches.
2. The prior knowledge represented by the pair wise constraints is useful for improving the performance of ISSCE.
3. ISSCE outperforms most conventional semi-supervised clustering ensemble approaches on a large number of datasets, especially on high dimensional datasets.

Proposed System Architecture:

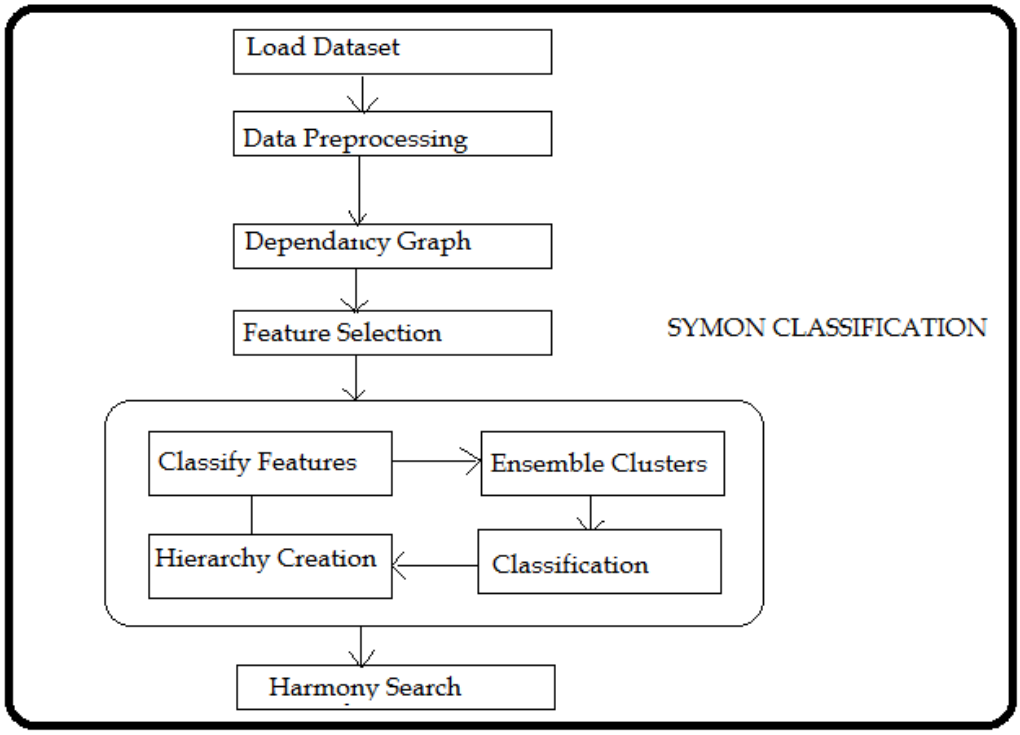

Fig. 1 System Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Work Flow Diagram:

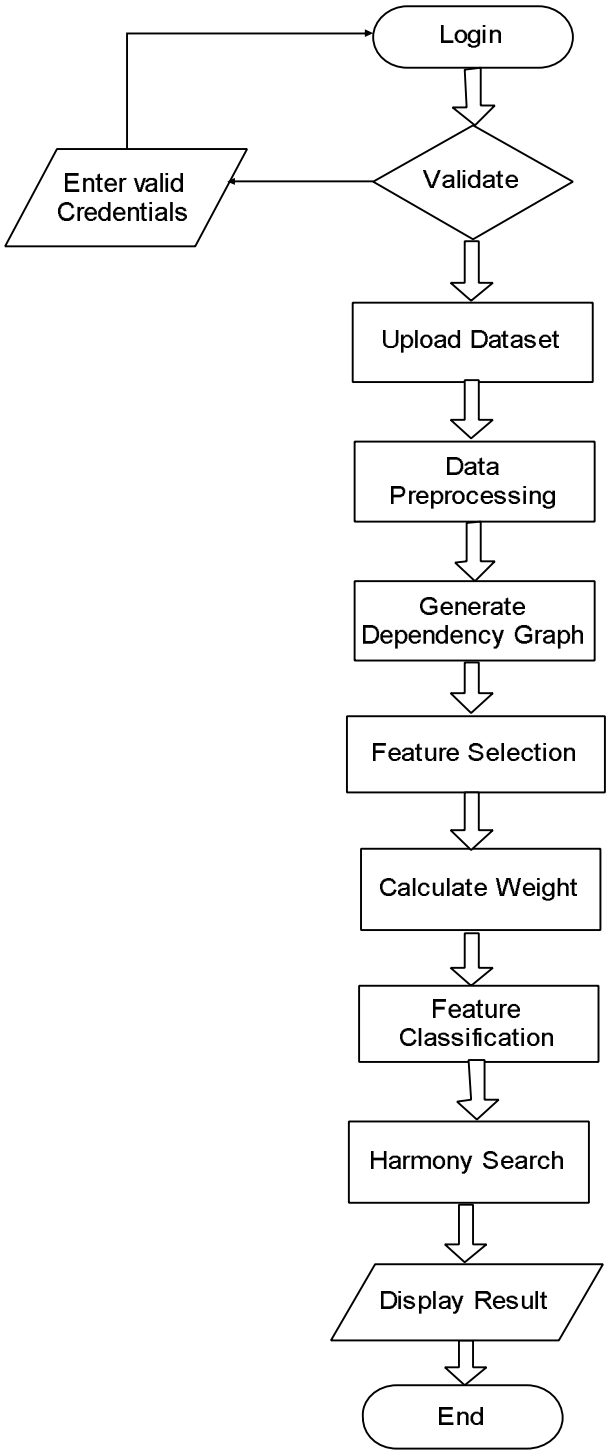

Fig. 2 Flow Chart

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Implementation Steps:

1. Upload Dataset: After login admin will get home page on that he/she will upload high dimensional imbalanced dataset.
2. Data Preprocessing: Data preprocessing step data transformation from string to numeric.
3. Generate Dependency Graph: Generate graph of attributes in the dataset to find correlation among the data items.
4. Symmetrical Uncertainty: Calculate weight of the given features using symmetrical uncertainty algorithm.
5. Feature Selection: Attribute based features selection process is implemented using SYMON. This process includes Area Under Cover and G- Mean algorithm to select features for data classification.

- Wrapper-based method: produce a more targeted feature subset.
- Top-k ranked features: selects weighted and ranked features.

6. Ensemble Clustering: Ensemble clustering means grouping similar data item from dataset according to their hierarchy.
7. Harmony Search: Relational classification is implemented throughout the data attributes classification by K – Nearest Neighbors for data.
8. Final step, logout successfully.

Algorithm1: SYMON

Input: F- Set of all features, C- Set of all class labels, NI- number of iterations, HMS- harmony memory size, HMCR- harmony memory consideration rate, PAR_{max}- maximum pitch adjustment rate, PAR_{min} minimum pitch adjustment rate.

Output: HM, optimized solution vectors in harmony memory

Process:

Step1: Start
Step2: w=CalculateSU (F, C)
Step3: Initialize ( )
Step4: **for** $t \leftarrow 1, \dots, NI$ **do**
Step5: **for** $f \in F$ **do**
Step6: $R \leftarrow random\ number$
Step7: **if** $R_f > HMCR$ **then**
Step8: Randomly select vector v_r from HM
Step9: $NHV|f| \leftarrow v_r|f|$
Step10: $R_p \leftarrow random\ number$
Step11: $P_f \leftarrow PAR(t)$ (Eq. (6))
Step12: **if** $P_f < R_p$ **then**
Step13: $NHV[f] \leftarrow \overline{NHV[f]}$
Step14: **else**
Step15: $\emptyset \leftarrow random\ number$
Step17: **if** $\emptyset > 0.5$ **then**
Step18: $NHV[f] \leftarrow 1$
Step19: **else**
Step20: $NHV[f] \leftarrow 0$
Step21: NHV = VectorTune (NHV, w, r, d);
Step22: **if** $f(NHV) > f(v)$ **then**
Step23: HM = HM — {v} U {NHV}
Step24: End

Algorithm2: Symmetric Uncertainty

CalculateSU ( )

Input: F- Set of all features, C- Set of all class labels

Output: w- Set of feature weights

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Process:
Step1: Start
Step2: **for** $f \in |F|$ **do**
Step3: **for** $c \in |C|$ **do**
Step4: Measure $S(f|c)$
Step5: Sum all the dependency values of f to class labels, $\sum_f S(f|c)$
Step6: $w(f)$ =calculate the final weight of f using Eq. (5)
Step7: End

Algorithm3: VectorTune ( )
Input: w- Set of feature weights, NHV- a feature vector, r- ripple factor, d- subset size
Output: Optimal feature subset vectors in harmony memory
Process:
Step1: Start
Step2: $F_s^w \leftarrow \{F_i^w, F_j^w, \dots\} \subset NHV$
Step3: $F_u^w \leftarrow F^w - F_s^w$
Step4: $F_s \leftarrow \{F_i, F_j, \dots\} \subset NHV$
Step5: $F_u \leftarrow F - F_s$
Step6: **if** $|F_s| = d$ **then**
Step7: F_s is changed by applying Ripple_Rem(r) and Ripple_Add(r)
Step8: **if** $|F_s| > d$ **then**
Step9: F_s is increased by applying Ripple_Add(r)
Step10: **if** $|F_s| < d$ **then**
Step11: F_s is decreased by applying Ripple_Rem(r)
Step12: tuned NHV
Step13: End

V. EXPERIMENTAL SET UP

The evaluation here will use a number of measures (classifier metrics, statistics and execution time) and different high-dimensional imbalanced datasets (microarray and imagery) for comparison against benchmark algorithms. For meaningful comparison, we draw upon the evaluation methods reported in similar works replicating the experiments using SVM as the underlying classifier and measuring the classifier performance using Area Under Curve (AUC), G-Mean (GM) and the Wilcoxon signed-rank sum. The results are promising and answer the following questions.

• What are the best empirical settings to ensure SYMON's optimal performance? SYMON's performance can be affected by the free parameters, so fine-tuning these parameters is crucial to ensuring the optimal performance of SYMON. Thus, we discuss how this near optimality can be achieved in Section 4.2.

• How comparable is SYMON's performance to existing state-of-theartfeature selection algorithms designed for high dimensional imbalanced class problems? This is clearly the key question that motivates the evaluation;hence we compared SYMON against similar works (Guyon et al., 2002; Yin et al., 2013; Maldonado et al., 2014) as discussed earlier in Sections 4.3.1 and 4.3.2. The related comparisons are made with SVM-RFE (Guyon et al., 2002), SVM-BFE (Maldonado et al., 2014) and Hellinger based feature selection algorithm (Yin et al., 2013).

• How effective is the performance of SYMON in comparison to SMOTE as one of the well-established baseline algorithms? This question investigates the performance of SYMON and compares against SMOTE (Chawla et al., 2004;Deepa and Punithavalli, 2011). To answer this question in Section 4.3.1 we integrate filter-based ranking algorithms of ReliefF (RLF)3and Principal Component Analysis (PCA) with SMOTE. The variations are called SMOTE-RLF and SMOTE-PCA.

• How robust is SYMON when presented with data sets possessing different levels of imbalance? Flowing from the first key question, SYMON's performance should be stable across a variety of data sets. We evaluate SYMON with different levels of class imbalance for different data sets and then compare the results against other works.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

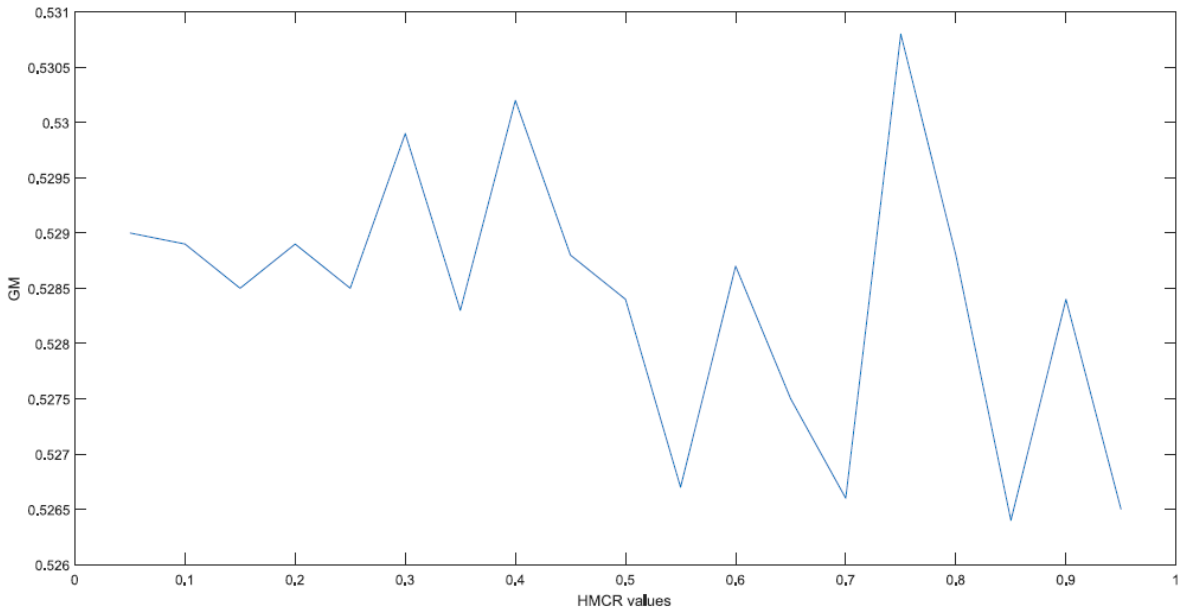

Fig. 3. HMCR value fine tuning with respect to Experiment.

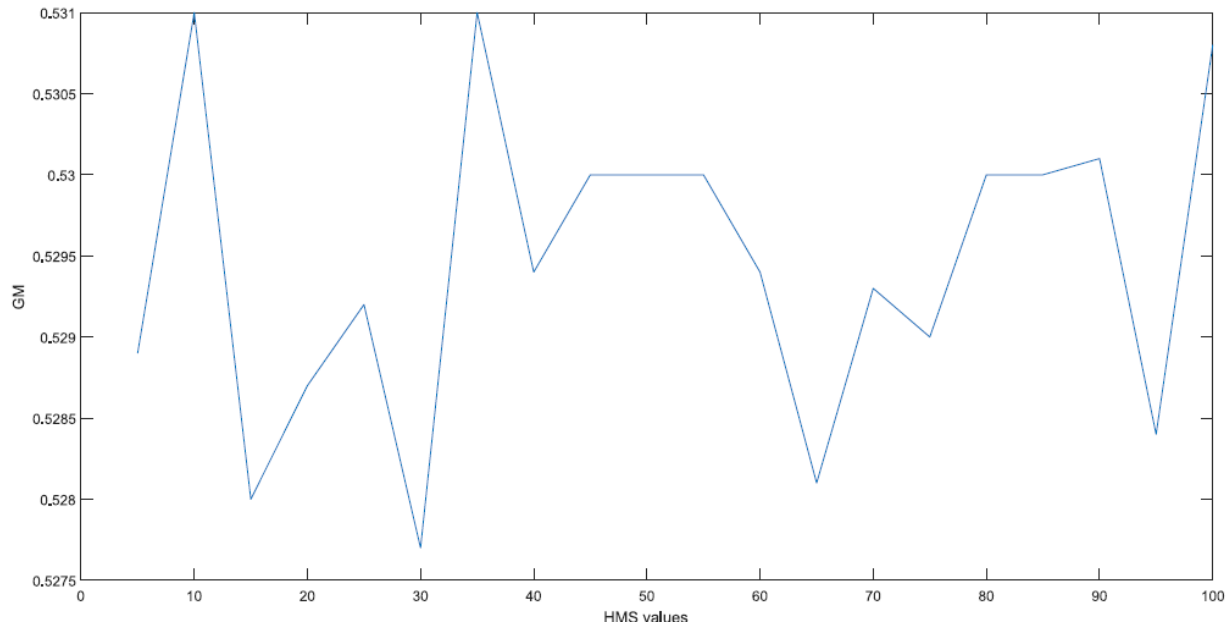

Fig. 4. HMS value fine tuning with respect to Experiment .

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

The best in all settings compared to the rest. In the other data sets:
CAR, LUG and BC, SYMON is either on-par or better across the various test settings. In the case of BC (d=F/5), while the G-Mean score is the same, SYMON uses smaller feature subsets than D-HELL to achieve this same score. This lower number of features has practical implications in terms of model interpretation. Finally, in the SRBCT data set, the performance of SYMON is similar to the other three under evaluation.

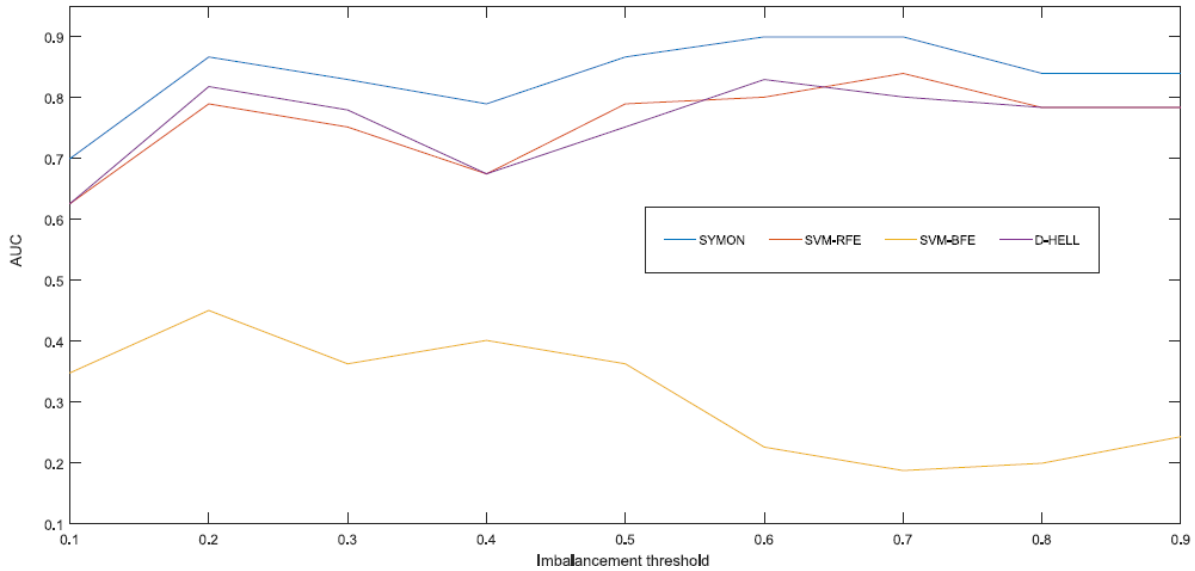

Fig. 5. Robustness of various algorithms to different rates of imbalancement based on AUC.

VI. CONCLUSION

Proposed system has introduced SYMON as a new feature selection algorithm for high dimensional imbalanced class datasets. SYMON has two stages algorithm, in first stage, feature weighting, measures the features' weights. In the second stage, known as feature selection, the top-k features are selected based on their weights. What distinguishes SYMON from similar works are (i) its capability in measuring the feature weight with respect to the dependency to class label(s) and (ii) dealing with the situation where different features have the same weight SYMON was empirically compared against the state-of-the-art and baseline algorithms and the results showed comparable or better performance over different high dimensional datasets. This performance can be attributed to its use of symmetrical uncertainty to weight features and the vector tuning operations embedded in the feature selection stage. On the limitations, SYMON has two that we will address for the future work. The first limitation is its high computational time. Even though we experimentally showed that SYMON can be improved in terms of execution time, by focusing on a proportion of the most significant features, a better solution is to explore a faster harmony search core to improve its runtime. The other limitation is to confine feature selection to a desired subset size (d). At the moment, the vector tuning operations are highly dependent on (d) and the ripple factor (r). Instead, a more flexible d will allow more optimal parts of the solution space to be discovered. This could be another avenue to improve SYMON's runtime performance.

Feature selection problems to support classifier ensemble reduction, by transforming ensemble predictions into training samples, and treating classifiers as features. Also, the global heuristic harmony search is used to select a reduced subset of such artificial features, while attempting to maximize the feature subset evaluation. The resulting technique is systematically evaluated using high dimensional and large sized benchmark datasets, showing a superior classification performance against both original, unreduced ensembles, and randomly formed subsets.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

REFERENCES

[1] Alibeigi, Mina, Hashemi, Sattar, Hamzeh, Ali, 2012. Dbfs: an effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data Knowl. Eng. 81, 67–103.
[2] Cerf, Loïc, Gay, Dominique, Selmaoui-Folcher, Nazha, Crémilleux, Bruno, Boulicaut, Jean-François, 2013. Parameter-free classification in multi-class imbalanced data sets. Data Knowl. Eng. 87, 109–129.
[3] Forsati, Rana, Mahdavi, Mehrdad, Shamsfard, Mehrnoush, Reza Meybodi, Mohammad, 2013. Efficient stochastic algorithms for document clustering. Inf. Sci. 220, 269–291.
[4] Forsati, Rana, Moayedikia, Alireza, Jensen, Richard, Shamsfard, Mehrnoush, Reza Meybodi, Mohammad, 2014. Enriched ant colony optimization and its application in feature selection. Neuro computing 142, 354–371.
[5] Forsati, Rana, Moayedikia, Alireza, Shamsfard, Mehrnoush, 2015. An effective web page recommender using binary data clustering. Inf. Retr. J. 18, 167–214.
[6] Huang, Yin-Fu, Lin, Sheng-Min, Wu, Huan-Yu, Li, Yu-Siou, 2014. Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. Data Knowl. Eng. 92, 60–76.
[7] Koprinska, Irena, Rana, Mashud, Agelidis, Vassilios G., 2015. Correlation and instance based feature selection for electricity load forecasting. Knowl.-Based Syst. 82, 29–40.
[8] Maldonado, Sebastián, Weber, Richard, Famili, Fazel, 2014. Feature selection for high dimensional class-imbalanced data sets using support vector machines. Inf. Sci. 286, 228–246.
[9] Nekkaa, Messaouda, Boughaci, Dalila, 2015. Hybrid harmony search combined with stochastic local search for feature selection. Neural Process. Lett., 1–22.
[10] Wang, Youwei, Liu, Yuanning, Feng, Lizhou, Zhu, Xiaodong, 2015. Novel feature selection method based on harmony search for email classification. Knowl.-Based Syst. 73, 311–323.