Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

JOSE MEASURE BASED HIGH DIMENSIONAL DATA CLUSTERING FOR REAL WORLD CONDITIONS

[1]M.SUGUNA, [2]Dr.S.PALANIAMMAL

[1]Assistant Professor, Department of Computer Science, Dr.G.R.Damodaran College of Science, Coimbatore

[2]Professor & Head, Department of Science and Humanities, Sri Krishna College of Technology, Coimbatore

E-mail: [1]suguna_a@yahoo.com , [2]splvlb@yahoo.com

ABSTRACT

The modern real world data has more number of dimensions with varying arguments. Different organization maintains customer or product information in different forms which is difficult to perform clustering. Each data point has different in size and properties, but has to be clustered in meaningful and efficient way to get some knowledge from that. Many strategies have been proposed for clustering high dimensional data but suffer with the problem of overlapping and retrieval efficiency. We propose a new measure called JOSE (Joint optimum similarity eccentric), which represents the distance between points in multiple forms of data. Jose measure, which shows the relationship between two data points at their junction points with eccentric data points. The Jose measure shows the closeness of particular data point with set of data points in a cluster. The proposed approach computes distance between data points with available attributes of data points with remaining data points from the cluster. The proposed approach has produced higher efficient clustering with negligible overlapping. We have used Enron data set for the evaluation of the proposed approach and the results shows that the proposed method has produced efficient result than others.

Keywords: *High dimensional data, Clustering, JOSE Measure.*

1. INTRODUCTION

The organization maintains various information about the customers and their product. There is a huge requirement of maintaining that information in an organized way, so that it could be used in an efficient way. The problem is with varying size of information and the size of data also with the number of attributes. There are numerous algorithms has been proposed earlier for clustering data but all or most of them for clustering data points with limited attributes. For example, the popular k-means algorithms have been suitable for clustering data points with limited or small number of attributes. Because in k-means the distance between data points are computed using Euclidean distance measure which is suitable for limited number of attributes.

Similarly the Bayesian approach also available for clustering data points where it uses certain rules and conditions to compute the distance between data points according to the probabilistic nature of data points. There are many other approaches available for clustering data points, but most of them suffer with the scalability in size and accuracy. Our requirement is to cluster data points which are in high dimension and more diverse in nature.

We have used real world data sets like Enron which is a data set which has information about YouTube and face book users. The real world data set can be used for clustering which has many number of dimensions, because a single user may have any number of contacts with various groups and each can be considered as a dimension. To identify or group similar interested users, the clustering approach can be used. The Enron data set has number of users and each has connected with other users of the same group and from other groups. The entry in the data set has a pair of numbers which shows that there is a connection between two users of the network. If a user is interested in data mining then the second user is also has interested in data mining. This kind of data set is useful in grouping similar interested users of the network to share information between them and so on.

The word high dimensional data means having number of attributes and also huge in size.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To cluster such data we need to design a metric to compute distance between data points. There are many number of algorithms have been introduced for clustering categorical data, each has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering. For example the feature-based approach transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels), the direct approach that finds the final partition through relabeling the base clustering results, graph-based algorithms that employ a graph partitioning methodology and the pair wise-similarity approach that makes use of co-occurrence relations between data points.

2. RELATED WORKS

A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data [1], where features are divided into clusters by using graph theoretic clustering methods. Then the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient maximum-spanning tree clustering method.

Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms [2], treat clusters as the dense regions compared to noise or border regions. Many momentous density based subspace clustering algorithms exist in the literature. Each of them is characterized by different characteristics caused by different assumptions, input parameters or by the use of different techniques etc. Hence it is quite unfeasible for the future developers to compare all these algorithms using one common scale. . This paper presents a review of various density based subspace clustering algorithms together with a comparative chart focusing on their distinguishing characteristics such as overlapping / non-overlapping, axis parallel / arbitrarily oriented and so on.

The Role of Hubness in Clustering High-Dimensional Data [3], show that hubness, i.e., the tendency of high-dimensional data to contain points (hubs) that frequently occur in k-nearest neighbor lists of other points, can be successfully exploited in clustering. The algorithm validated the hypothesis by demonstrating that hubness is a good measure of point centrality within a high-dimensional data cluster. It also proposes several hubness-based clustering algorithms, showing that major hubs can be used effectively as cluster prototypes or as guides during the search for centroid-based cluster configurations.

Detection of Arbitrarily Oriented Synchronized Clusters in High-Dimensional Data [4], propose ORSC (Arbitrarily ORiented Synchronized Clusters), a novel effective and efficient method to subspace clustering inspired by synchronization Synchronization is a basic phenomenon prevalent in nature, capable of controlling even highly complex processes such as opinion formation in a group. Control of complex processes is achieved by simple operations based on interactions between objects. Relying on the interaction model for synchronization, the proposed approach ORSC naturally detects correlation clusters in arbitrarily oriented subspaces, including arbitrarily shaped non-linear correlation clusters. ORSC approach is robust against noise points and outliers.

Clustering Algorithm for High Dimensional Data Stream over Sliding Windows [5], proposes an effective clustering algorithm referred as HSWStream for high dimensional data stream over sliding windows. This algorithm handles the high dimensional problem with projected clustering technique. It deals with the in-cluster evolution with exponential histogram of cluster feature called EHCF and eliminates the influence of old points with the fading temporal cluster features. A fast computational method is employed to maintain the exponential histogram cluster feature. The algorithm produces high quality clusters with less memory usage.

PROCLUS [6], is an efficient high dimensional clustering algorithm; consist of significant issues like inconsistency in results and expert supervised subspaces. MPROCLUS: a modified PROCLUS algorithm is proposed, aimed

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

at improving the running time and consistency as well as the unsupervised selection of the parameter like, average number of dimensions. The promising and consistent results of MPROCLUS has open the sky wide open for further research for usage of MPROCLUS in stream Data Mining.

A Relevant Clustering Algorithm for High- Dimensional Data [7], is used for finding the subset of features. A Relevant clustering algorithm renders efficiency and effectiveness to find the subset of features. Relevant clustering algorithm work can be done in three steps. First step involves elimination of irrelevant features from the dataset; the relevant features are selected by the features having the value greater than the predefined threshold. In the second step, selected relevant features are used to generate the graph, divide the features using graph theoretic method, and then clusters are formed by using Minimum Spanning Tree. In the third step, the subsets features that are more related to the target class are selected.

Non-parametric detection of meaningless distances in high dimensional data [8], quantify the phenomenon of high dimensional clustering more precisely, for the possibly high but finite dimensional case in a distribution-free manner, by bounding the tails of the probability that distances become meaningless. As an application, they show how this can be used to assess the concentration of a given distance function in some unknown data distribution solely on the basis of an available data sample from it. This can be used to test and detect problematic cases more rigorously than it is currently possible. The paper demonstrates the working of this approach on both synthetic data and ten real-world data sets from different domains.

Fast agglomerative clustering using information of k-nearest neighbors [9], develop a method to lower the computational complexity of pair-wise nearest neighbor (PNN) algorithm. This approach determines a set of candidate clusters being updated after each cluster merge. If the updating process is required for some of these clusters, *k*-nearest neighbors are found for them. The number of distance calculations for our method is $O(N^2)$, where N is the number of data points. To further reduce the computational complexity of the proposed algorithm, some available fast search approaches are used. The algorithm was compared with FPNN and PMLFPNN and inferred that the computing time and number of distance calculations are reduced significantly by a factor about 26.8 and 3.8 respectively for the same dataset.

Hubness-Based Fuzzy Measures for High-Dimensional k-Nearest Neighbor Classification [10] proposed several fuzzy measures for *k*-nearest neighbor classification, all based on hubness measure, which expresses the fuzziness of elements appearing in *k*-neighborhoods of other points. Experimental evaluation on real data from the UCI repository and the image domain suggests that the fuzzy approach provides a useful measure of confidence in the predicted labels, resulting in improvement over the crisp weighted method, as well the standard *k*NN classifier.

A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian kNN [11], present a new probabilistic approach to kNN classification, naive hubness Bayesian k-nearest neighbor (NHBNN), which employs hubness for computing class likelihood estimates. Experiments showed that NHBNN compares favorably to different variants of the kNN classifier, including probabilistic kNN (PNN) which is often used as an underlying probabilistic framework for NN classification, signifying that NHBNN is a promising alternative framework for developing probabilistic NN algorithms.

Hubs in space: Popular nearest neighbors in high-dimensional data [12], explores a new aspect of the dimensionality curse, referred to as hubness, which affects the distribution of *k*-occurrences, the number of times a point appears among the *k* nearest neighbors of other points in a data set. Through theoretical and empirical analysis involving synthetic and real data sets the algorithm show that under commonly used assumptions this distribution becomes considerably skewed as dimensionality increases, causing the emergence of *hubs*, that is, points with very high *k*-occurrences which effectively represent “popular” nearest neighbors.

Problem Formulation:

The hubness [10] and density-based clustering [2] performs grouping of data points according to the weightage of data points around the centre of cluster which spoils the subspace clustering because, the presence of other data points in the cluster are ignored. The fast clustering [1], and probabilistic knn [11] approaches follows

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

probabilistic clustering which generates clusters only based on probability values and struggles identifying the exact subspace of the cluster, because the probability is produced based on number of data points present in the subspace of the cluster. Fast agglomerative clustering [9], performs clustering in iterative manner which increases the time complexity and needs merging of clustering again and again to ensure the clustering efficiency. The sliding windows [5], Proclus [6], relevant clustering [7] has been discussed to generate clustering in tree based approaches which uses graph and other measures to produce clusters, and needs more number of iterations to be performed to produce final clusters. All the above discussed approaches need more time and require more number of iterations to be performed and produces inefficient clusters. This motivates us to identify some other efficient measure to perform clustering and to identify closeness between data points with high dimensions.

3. PROPOSED METHOD

We propose a new strategic approach to cluster high dimensional data points. In the proposed approach the distance between data points are computed using Jose Measure. Based on the compute Jose measure the data points are assigned with labels.

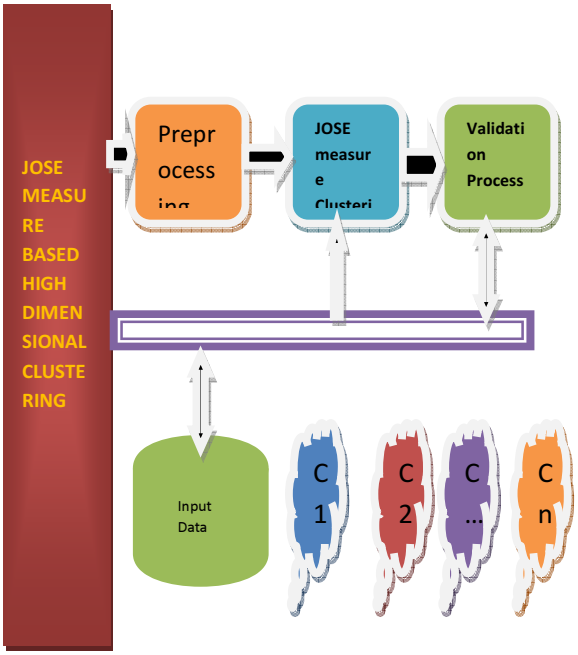


Figure1: Proposed System Architecture.

Given data set Ds with N data points, and C category, the proposed method clusters the data points according to the distance similarity value. The distance similarity value is computed using JOSE measure, and the similarity of data points are computed at all attributes of data points. The proposed approach has following steps:

3.1 Preprocessing:

Generally the data points with missing values are identified. The identified missing values are assigned with averaging factor of other data points. Assigning averaging data points increases the cluster efficiency and avoid the data point been rejected from clustering. The computation of averaging factor is performed as follows: For each missing attribute, from all the data points from all the clusters we compute the average value for that particular attribute and that value will be assigned to the data point.

Preprocessing Algorithm:

Input: Raw Data set Ds.

Output: Preprocessed data set.

Step1: Initialize attribute set As.

Step2: for each data point D_i from Ds

Extract attributes A = $\sum Ai \in Di$

Add attribute to As where $A_i \notin As$

As = $\sum Ai(As) + A(Ai) \notin As$

end.

Step3: for each data point D_i from Ds

if $\int_{i=1}^{N} if\ Di \notin Ai$ then

Remove the data point from Ds.

end.

end.

Step 4: stop.

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Jose Measure Based Clustering

In this approach the attributes of data points are identified and for each identified attribute we compute the diverse factor which represents the mean value of that particular attribute. Similarly we compute the diversity factor for each of the attribute identified. Then from the entire diversity factor computed, we compute the overall diversity factor which shows the closeness of data point with the cluster. Similarly this will be performed for each of the cluster and finally a single cluster will be selected which has more closure value. The jose measure, is computed for both numerical and non numeric values, in case of non numeric values , the proposed method uses enum values and rules.

Input: Preprocessed Data Set Ds.

Output: Cluster Sets.

1. Identify set of unique attributes set As.
2. for each attribute A_i from As

compute diversity factor Df = $\int_1^N \Sigma(Ai - Aj)/N$

end

3. compute jose measure Jm = Ø(As)*(ΣDf/N)
4. select top Jm from all cluster.
5. Assign label to data point and return clusters.

3.3 Validation Process

The generated clusters are validated for its correctness and to measure the clustering accuracy in this stage. We select set of random data points from each cluster generated and for each selected data point we compute the JOSE measure. According to the computed JOSE measure, a new cluster label is identified and measured for the accuracy of clustering based on early labels assigned and new labels produced.

4. RESULTS AND DISCUSSION

The Jose measure based high dimensional clustering has been evaluated for its efficiency using different data sets. Each data set has large number of records with many numbers of attributes. We have evaluated the proposed method by splitting 70 percent as training set and 20 percent as test set. The proposed method produced efficient results with various data sets. We have evaluated the algorithm with the following data sets.

Dataset	Number of Data Points (N)	Attributes (d)	Attribute Values (AA)	Classes (K)
zoo	101	16	36	7
20 news group	1000	6084	12,168	2
Mushroom	8124	22	117	2
KddCup99	1,00,000	42	139	20
Enron	32,378	164	245	2500

TABLE 1: Description of Data Sets

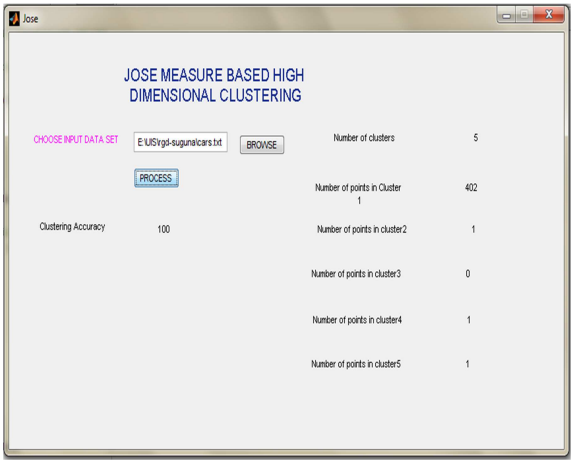
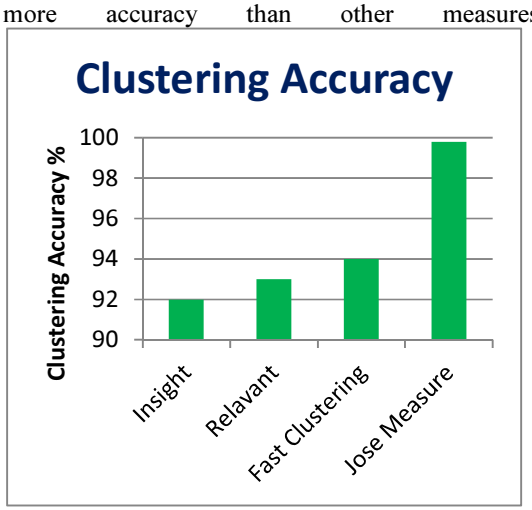


Figure2: Snapshot Of Proposed Approach

The figure2 shows the result produced by the proposed method, and it shows that the proposed method has produced clustering with more accuracy than other measures.



Graph1: Comparison Of Clustering Accuracy.

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The graph1, shows the clustering accuracy produced by various methods. It shows that the proposed approach has produced more accurate clusters than other methods. The clustering accuracy of the proposed method has been evaluated by selecting random data points from each cluster which are already indexed. From the selected data points we compute the same JOSE measure between the data points of each cluster C_i, and a single cluster is selected for indexing.

The accuracy of the proposed approach has been evaluated as follows:

$$CA = \frac{NTP + NTN}{NTP + NTN + NFP + NFN} \quad \text{---------- (1).}$$

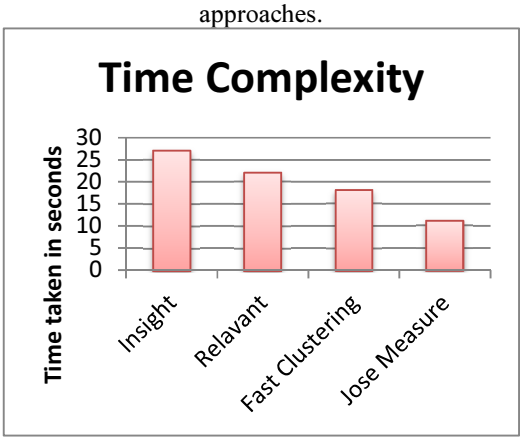
NTP– Number of classification of true positive

NTN- Number of classification of true negative

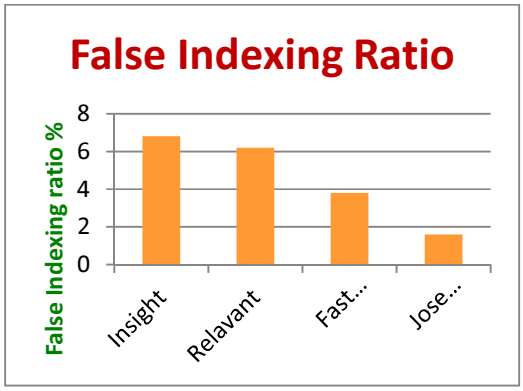
NFP- Number of classification of false as positive

NFN – Number of classification of false negative.

The graph2 shows the classification time which is claimed by the different methods to identify or cluster set of data points given as testing. It shows that the proposed JOSE measure based approach has claimed less time than other approaches.



Graph 2: Shows The Time Complexity Of The Proposed System.



Graph3: Shows False Indexing Ratio Of Different Algorithms

The graph3 shows the ratio of false indexing produced by different algorithms for same size of training and testing set. It shows clearly that the proposed approach has produced efficient results with least false indexing ratio.

Similarly the false indexing ratio is computed as follows:

False indexing ration FIR = $\frac{\text{Number of data points indexed incorrectly}}{\text{Total number of points selected.}}$

We have evaluated the generated clustering and effectiveness of the proposed approach with various other measures also.

Table2: Efficiency Parameter Analysis.

Precision	Recall	Entropy	Scatter
0.92	0.97	0.92	0.86

The table 2, shows the cluster efficiency of the proposed approach in producing clusters and list out the parameters of the clustering efficiency by which it is gauged.

The precision value of the clustering approach is calculated as follows:

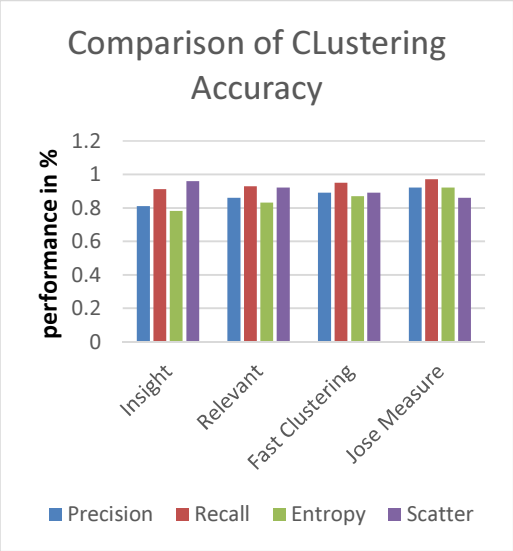
Precision $= \frac{NTP}{NTP + NFP}$, shows the relevancy of document retrieved.

Recall $= \frac{NTP}{NTP + NFN}$, shows the query relevancy of document retrieved

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Entropy = $\sum_{i=1}^{N} NFP * Log(NTP)$ - the entropy shows the purity of cluster.

Scatter – shows the closeness of data points in the cluster towards the center of cluster.



Graph4: Comparison Of Clustering Accuracy Of Different Methods.

The Graph4, shows the comparison of different parameters of clustering accuracy like precision, recall, entropy and scatter values computed for each algorithm and it shows that the proposed JOSE Measure based clustering has produced efficient results in all the parameters of efficiency.

The proposed method has been evaluated for its efficiency with various parameters of the clustering accuracy and it shows that the proposed approach has produced effective results in all the factors and produced effective clustering.

5. CONCLUSION

We proposed a novel clustering measure called Joint Optimum Similarity Eccentric (JOSE) for clustering high dimensional data sets. The proposed approach performs preprocessing of data set in order to remove the noisy data points which are incomplete or missing values. The preprocessed data points are randomly assigned to set of pre assigned clusters and then based on JOSE measure other points are clustered. The proposed method has been validated by randomly picking data points from clustered data points and measured for its accuracy. The proposed approach has produced efficient results and has reduced false indexing ratio. The time complexity of the proposed approach is comparatively less than previous methods.

REFERENCES:

[1] Q. Song, J. Ni, G. Wang, **A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data,** IEEE Transactions on Knowledge and Data Engineering, 2011.

[2] Sunita Jahirabadkar and Parag Kulkarni. Article: Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms. *International Journal of Computer Applications* 63(20), February 2013.pp:29-35

[3] Nenad Tomasev, The Role of Hubness in Clustering High-Dimensional Data , ieee transactions on knowledge and data engineering, revised january 2013.

[4] Junming Shao, Detection of Arbitrarily Oriented Synchronized Clusters in High-Dimensional Data, International conference on data mining,2011, pp: 607-616.

[5] Weiguo Liu, Clustering Algorithm for High Dimensional Data Stream over Sliding Windows, Trust, Security and Privacy in Computing and Communications (TrustCom),, 2011, pp:1537-1542.

[6] R G Mehta, N J Mistry and M Raghuwanshi. Article: Towards Unsupervised and Consistent High Dimensional Data Clustering. International Journal of Computer Applications 87(2):, February 2014, pp:40-44.

[7] Bini Tofflin.R1 , A Relevant Clustering Algorithm for High- Dimensional Data , International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014

[8] A. Kab ´an, "Non-parametric detection of meaningless distances in high dimensional data," Statistics and Computing, vol. 22, no. 2, 2012, pp. 375–385.

[9] C.-T. Chang, J. Z. C. Lai, and M. D. Jeng, "Fast agglomerative clustering using information of k-nearest neighbors," Pattern Recognition, vol. 43, no. 12, 2010, pp. 3958–3968.

[10] N. Toma˘sev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "Hubness-based fuzzy measures for high-dimensional knearest neighbor classification," in Proc. 7th Int.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Conf. on Machine Learning and Data Mining (MLDM), 2011, pp. 16–30.

[11] Nenad Tomsay "A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian kNN," in Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM), 2011,pp. 2173–2176.

[12] M Radovanović , "Hubs in space: Popular nearest neighbors in high dimensional data," Journal of Machine Learning Research, vol. 11, 2010, pp. 2487–2531,.

[13] N. Tomaˇsev and D. Mladeni´c, "Nearest neighbor voting in high dimensional data: Learning from past occurrences," Computer Science and Information Systems, vol. 9, no. 2, 2012, pp. 691–712,.

[14] N. Tomaˇsev, R. Brehar, D. Mladeni´c, and S. Nedevschi, "The influence of hubness on nearest-neighbor methods in object recognition," in Proc. 7th IEEE Int. Conf. on Intelligent Computer Communication and Processing (ICCP), 2011, pp. 367–374.

[15] K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and effective instance selection for time-series classification," in Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Part II, 2011, pp. 149–160.