Volume 2, Issue 3 March 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

IRJIF IMPACT FACTOR: 3.01

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi, Urdu all research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.01

HR RECRUITMENT AND SELECTION PROCESS FOR MANAGEMENT USING DATA OREINTED MULTI-INDEX HASHING

JAMDAR BHAGYASHREE, BAIRAGI HEMLATA, WARINGE PRAMILA & PROF. LOMESH AHIRE*

*Assistant professor, Department of Inforamtion Technology,
NMVPM's, Nutan Maharashtra Institute of Engineering and Technology, Talegaon Dabhade, Pune-410507

ABSTRACT –***Multi Index hashing (MIH) is method which divides long codes into substrings and hence builds multiple hash tables. But MIH method is based on assumptions that substrings are uniformly distributed. However, due to non uniform distribution the efficiency decreases. To overcome this factor, Data Oriented Multi Index Hashing (DOMIH) is proposed. DOMIH method corelates bits of code and using Naïve Bayes classifier, multiple hash tables are builds with uniforms distributions of codes.Using Fast Condensed Nearest Neighbour ,the compilation of cluster on the basis of statiscal properties , the retrivation of data can be obtained quickly as compared to KNN Method.Also by usage of Support Vector Machine (SVM) misclassification of traning data points near to decision boundry of bianary classifier are avoided. Using FCNN rule,the search and retrivation efficiency can be increase upto 25% - 27% .***

I. INTRODUCTION

Nearest neighbour is simple classifier which treats each instants in training set as exemplar itv expects that traing instances shoud not have , two very similar instances belonging to different classes nearest neighbour classifier has low bias and high variance performance of nearest neighbour classifier is not good if input inastant space has many features .i.e. curse of dimensionality.This model uses spaces calculated in cartesian coordinate system number of dimensions in cartesian coordinate system is equal to no of features no of dimension can be reduce by using sofisticated method called as feature selection and principle component analysis. The simplest method of nearest neighbour classification is FCNN classification.

Although binary codes can be directly used as indices of the hash tables, correlations between the bits may lead to non-uniform codes distribution and reduce the performance of the hash table. . Experiments conducted on a huge amount of binary codes extracted from the UK Bench dataset show that our method can achieve significant acceleration in searching speed for large scale dataset.DOMIH method has three major differences between MIH. Firstly, training set is build to compute the correlations using Navie bayes between bits of the codes and learn an adaptive projection vector for each substring. Then, instead of using binary substrings as direct indices into a hash table, project the substrings with corresponding projection vectors to generate new indices. With adaptive projection, the indices in each hash table of DOMIH are more near uniformly distributed than that in MIH. So it handles the "non-uniform distribution" problem to some extent. By assigning different bit level weights to different bits, the returned binary codes are ranked at a finer-grained binary code level.

By applying FCNN, condensed subsets with instances close to the decision boundary are obtained. The main objective of this approach is to improve the performance of the classification by boosting the quality of the training-set. The experimental results on several standard classification databases illustrated the power of the proposed method. In comparison to previous approaches which select prototypes randomly, training with High Power Prototype performs better in terms of classification accuracy.

II. EXISITNG METHODOLOGY

Previous HR Management system which handles the process of recruitment of fresherspossed of number of paper based work.Maintainence of these paper was complex and the track of the procedure was proved to be very difficult.Hence, new system was developed which will handle all the acivities including the database and the track record of the procedure.The proposed system based on the concept of machine learning used SVM techniques to analyse the datasets provided and accordingly will work to present a required output.Existing system used MIH method to retrieve the data from the database which consumed a large period of time thus drcresing the efficiency,However, MIH is based on the assumption that the codes in the dataset are distributed uniformly . Actually, the codes are not uniformly distributed, especially for themultimedia data . Besides, there are a lot of data items sharing the same Hamming space to a query andthe ranking of these data items is ambiguous.

So MIH has three short comings:
1) For the candidate buckets inthe multi-hash tables, ifthey have too many items, then there are too many candidate codes need checking for validity. So it costsmuch time for candidate codes checking.
2) For the candidate buckets inthe multi-hash tables, if they have too few items, then in search process the value ofneeds set to be large enough to ensure thatenough exact near neighbors are found. So it costs much time for index lookups.
3) For the applications, such as image retrieval and computervision,where ranking of data items is important,

MIH cannot distinguish the binary code sharing the same Hamming space to the query. Also MIH method consider only the problem of searching for keys, and thus cannot capture the relevance of the documents stored in the system.
This common problem with existing traditional distributed hash table is done because they usually ignore the information retrieval algorithms, and thereby rely on keyword based searches.

III. SCOPE

Data reduction is one of the most important problems for work with huge data sets. Usually, only some of the data points are needed for accurate classification. Those data are called the prototypes and can be found as follows:

Select the class-outliers, that is, training data that are classified incorrectly by k-NN (for a given k)

Separate the rest of the data into two sets:

(i) the prototypes that are used for the classification decisions and

(ii) the absorbed points that can be correctly classified by k-NN using prototypes. The absorbed points can then be removed from the training set.

Fast Condensed nearest neighbor (FCNN, the Hart algorithm) is an algorithm designed to reduce the data set for k-NN classification. It selects

the set of prototypes U from the training data, such that 1NN with U can classify the examples almost as accurately as 1NN does with the whole data set. Three types of points: prototypes, class-outliers, and absorbed points.

Given a training set X, CNN works iteratively:

1. Scan all elements of X, looking for an element x whose nearest samples from U has a different label than x.
2. Remove x from X and add it to U
3. Repeat the scan until no more samples are added to U.

Use U instead of X for classification. It is efficient to scan the training examples in order of decreasing border ratio.The border ratio of a training example x is defined as

$$a(x) = \frac{\left|\left|x' - y\right|\right|}{||x - y||}$$

where ||x-y|| is the space to the closest example y having a different color than x, and ||x'-y|| is the space from y to its closest example x' with the same label as x.

IV. MOTIVATION

A novel algorithm for the calculation of a training set of a regularly occuring subset for the NN rule is presented. The algorithm, i.eFast CNN rule works as follows. First, the constant subset S is initialized to the centers 1 of the classes contained in the training set T. Then, during each variation of the algorithm, for each point p in S,a point q of T belonging to the Voronoi cell of p,2 but having a different class label, is selected and added to S. The algorithm stops when no further point can beadded to S, i.e. when T is correctly classified using S.Despite being quite simple, the FCNN rule has somedesirable properties. Indeed, it is order not dependent,has subquadratic time complexity, requires few variations to converge, and it is likely to select points very close to the decision location.

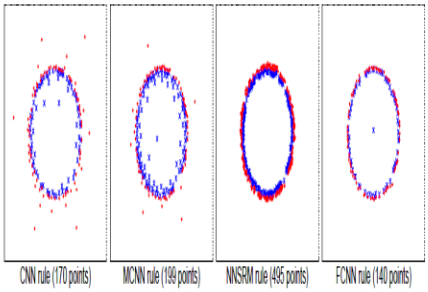


Figure 1. Example of training set consistent subsets computed by the CNN, MCNN, NNSRM, and FCNN rules.

For example, Figure 1 compares the constant subsets computed by the CNN, MCNN, NNSRM, and FCNN rules on a training set composed by 9,000 points uniformly distributed into the unit square and separation in two classes by a circle of diameter 0.5. As already noted, training set reduction algorithms can be characterized by their storage reduction, classification speed increase, generalization accuracy, noise tolerance, and learning speed. Among these criteria, the learning speed one is usually to failed. But, in order to be practicable on large training sets or in knowledge discovery applications requiring a learning step in their cycle, the method should exhibit good learning behavior.

The contribution of work can be summarized as follows.

1. Fast condensed nearest neighbor (FCNN) rule, is proposed for the calculation of a training set constant subset for the nearest neighbor rule.

2.The Fast condensed nearest neighbor (FCNN) rule has

1 Given a set S of points having the same class label,the center of S is the point of S which is nearest to the geometrical center of S.

2 The Voronoi cell of point p ∈ S is the set of all points that are closer to p than to any other point in S.worst case sub-quadratic time requirements, and de-scribe an implementation exploiting the triangular in-equality that sensibly reduces the worst case calculational cost.

3 comparing the FCNN rule withstate of the art competence preservation algorithmson large and high dimensional training sets, showingthat the Fast condensed nearest neighbor rule out performs existing methods interns of learning speed and learning scaling behavior,and in terms of size of the model, while it guarantee a comparable future evention correctness.

V. ISSUES IN EXISTING SYSTEM

The main disadvantage of is that the algorithm must calculate the space and sort all the training data at each future evention, which can be slow if there are a large number of training examples. Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being strong to clamorous data. Further, changing *K* can change the resulting future evented class label.The K-nearest neighbor classification rule (KNN) proposed by T. M. Cover and P. E. Hart, is a powerful classification method that allows an almost in fallible separation of an unknown sample through a set of training samples. It is widely used in pattern recognition text categorization, object recognition and event recognition applications.An prevent consequence of large sets of prototypes is the calculational time implied by this research problem. The databases,used in some areas such as intrusion detection,are constantly and dynamically updated.

This constitutes one of the main in conveniences of the nearest neighbor classification rule. Another important in convenience comes from the fact that the training prototypes can contain clamorous or mislabeled models that may affect the results and distort them.The scientific community has tackled these problems and proposed a selection of prototypes which could modify an initial set of prototypes by reducing its size in order to improve the separation performance.

VI. PROPOSED SYSTEM

Binary present again for large scale nearest neighbor search received more and more concern recently. Although binary codes can be directly used as indices of the hash tables, comparable between the bits may lead to non-uniform codes distribution and reduce the performance of the hashtable. In this paper, we propose a data driven multi-index hashing method for exact nearest neighbour search in Hamming space. By exploring a quantity calculated from the data in a sample properties of the dataset, we can separate the correlated bits into different segments during the procedure of building multiple hash tables, and thus make binary codes distributed as uniformly as possible in each hash table. Experiments conducted on a large amount of binary codes extracted from the UK Bench dataset show that our method can achieve significant acceleration in searching speed for large scale dataset.This Application is developed for Recruiting Process; we have divided this application in three STAGEs.

STAGE I

Functional Head receive the MPR forms regarding application for various job profiles. Functional Head send these MPR forms to the Head of Department or Managing Director for the approval
of MPR forms.MD or HOD approves the MPR form for specified job designation and sends back to the HR for further process. In case the forms are rejected, they are returned back to the Functional Head. When the accepted forms are handed over to HR, then department will assign reasonable salary to the specified job designation. This process is done by comparing the salaries between the employee staff. If a person needs to apply for other job designation, he can apply for the post internally using IJP (Internal Job Posting). If the application is accepted then the job position gets closed and the selected employee gets transferred.If any candidate is unavailable for position then HR needs to search for Profiles which are externally referred by using External Job Posting or internal resume database.If no candidate is referred through internal database then various process such as Searching profiles through Job Portals, Hiring Consultants, Campus Placement or advertisement.

STAGE II:-

Various Resume from different sources are received by HR .These Resumes are internally screened and selected Resume are further sent for next approval.canceled Resume are sent out and care is taken that they remain in the database for further references. first interviews are arranged by HR coordinators along with Hiring Manager for selection of candidates.Those candidates who successfully clear the preliminary interview need to fill-up the Interview Assessment Form.second interview for selected candidates are examined by HOD and final candidates are shortlisted and recommended to HR.If any candidate is selected for higher position, interview is supervised by Management Director and selected candidate is again recommended to Head Of Department. Rejected candidate by MD are removed from database

STAGE III:-

Selected candidates undergo the final interview conducted by HR and salaries are negotiated. If any candidate decline the salary offer then those candidates are hold on a standby and again the process of interview by HR are taken, provided vacancies are available. If any candidate accept the
offer, the position for particular designation gets closed and HR issues the offer letter and provide approximate date of joining. HR informs the concerned Functional head regarding the joining date of particular candidate for specified job designation.HR send mail to IT and concern Department head for Email id creation, allotment and procurement of computer for selected candidate appointed for particular post. On the date of joining, candidate undergoes the procedure of medical check-up and report gets submitted to Functional Head. Now all the procedure of joining, detailed appointment and Buddy letter are issued by the Functional Head.

VII. RELATED WORK

The algorithm for the Fast Condensed nearest neighbour rule can be described as follows:
1. Copy the first sample (T1) in the training set to the minimum subset (M1).
2) Classify the next sample in the training set using the samples in the minimal subset (M). 3) If the classification is correct, repeat step 2 until all the

samples are classified correctly. If all samples are classified correctly, STOP 4. If the classification is incorrect, copy the incorrectly classified sample Ti to the minimum subset M and restart the classification process from 2.

Online Learning method:

On-line learning works in iteration. If there are T training data points then online learning algorithm follows 'T' number of iteration.In some iteration 't',t^{th} data point feature vector 'X_t' is given as input. Current learning model say 'h_t' is used to future event $\widehat{Y}_t$ the value associated with X_t as $\widehat{Y}_t$=$h_t(X_t)$. Then actual value of Y_t associated with X_t in training data point is provided to learning algorithm.If future evented value $\widehat{Y}_t$ and actual Y_t associated with X_t are same,then no changes are made to the learning model (h_t).
If future evented value $\widehat{Y}_t$ and the actual value Y_t associated with X_t are not same then learning model h_t is adjusted or refined to h_{t+1} .

Algorithm:
1.Initialization step : select hypothesis h_0.
2.For each training data point t=1,2,……..T
{
3.Observe t^{th} training data input feature vector X_t
4. Apply current Learning model h_t to X_t to future event $\widehat{Y}_t$ as $\widehat{Y}_t$=$h_t(X_t)$
5. Observe correct value Y_t associated with X_t from input training data
6. Check for error.
If $\widehat{Y}_t = Y_t$ i.e $h_t(X_t) = Y_t$
Then $h_{t+1} = h_t$
Else
//in case of error
h_{t+1}=updated h_t which improves error for X_t

VIII. METHODOLOGY

FCNN Method:

Fast retrivation of data from the dataset is essentiall andis done by using Fast condensed neartest neighbour method.The method can br decribesd as follows.

$S = \emptyset;$

$\Delta S = \text{Centers(T)};$

$\text{while}(\Delta S \neq \emptyset)\{S = S \cup \Delta S\};$

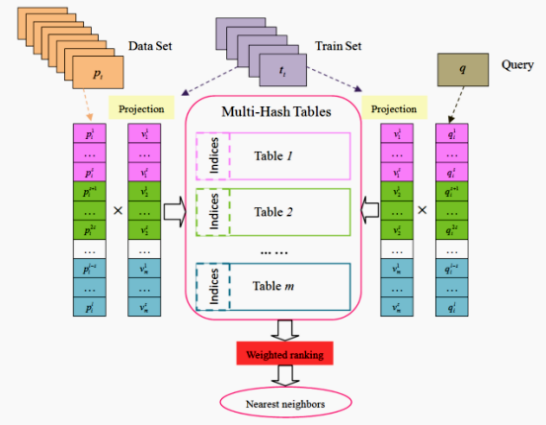
$\Delta S = \emptyset;$

For each $(p \in S)$

$\Delta S = \Delta S \cup \{\text{rep}(p, \text{Voren}(p, S, T))\};$

}Return(S);

IX. SYSTEM ARCHITECTURE



Data oriented mult index hashing techniqueposses a simple architecture which consist of data sets which are in non unifprm distribution. Using support vector machine, the random datasets can be trained to result in a perfect uniform distributed training set. This

training set is projected with the help of PCA(Principle Component analysis) in vector indices.
This indices form to be the serial communication between the hash tables indices and the projection vectors of training set.When a particular query is fired to retirve the data from the dataset these projection vectors are directed to the indices of the hash tables and the correlation between the matrices is maintained.Along with the projection the hashtables are assigned and ranked in a proper sequence ahd hence there is no random distributuion of hashtables which result in increased search efficiency.

The training set build is used to compute the correlations between bits of codes and machine learning algorithm helps to learn the adaptive projection vectors for each substring. the binary substrings are directed to the indices of hashtables and generate new indices .

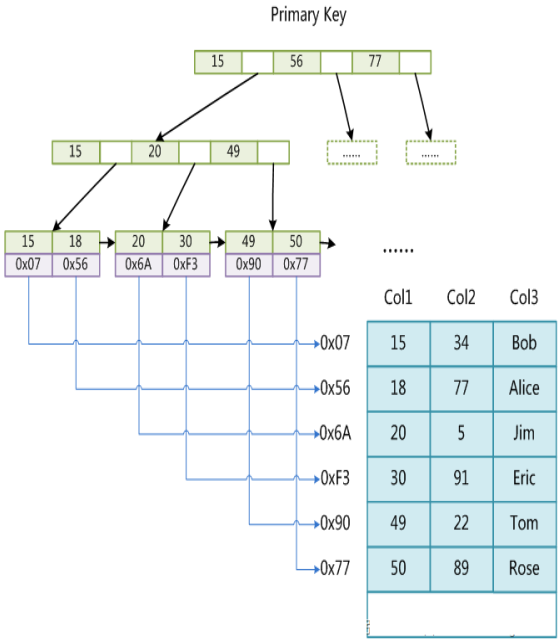
With adaptive projection,the indices (data items) in each hash table of DOMIH are more near uniformly distributed than that in MIH.The proposed system handles the “non-uniform distribution” problem not fully but to a certain limit. Finally, a ranking method for the binary codes with the covariance matrix is used to rank the hash tables in a sequence . By assigning different bit level weights to different bits, the returned binary codes are rank data finergrained binary code level.

The bits of code which correlates the projection vector with the indices of the hash tables are build on the basisi of Navies bayer classifeir..This algorithm is used to classfiy an observation in more than or equal to 2 classes.Navies bayes classifier also finds probability at which an observation may belong to certain class.

Navies bayes classifier can be used for classification of categorial data.The uniform distribution os dataset are categorised on the basis of nearest cluster and ranking method so as to output a proper sequence of data.for particular retrivation of data (resume of candidates) are categorised on the basis of knowledge skills. For the retrivation process and sequence output Navies bayes classifier can be implemented.

Support vector machines is used for to classification of instances in 2 classes say +1,-1.Support vector machine algorithm sets boundary to the dataset to form the clusters on the basis of nearest neighbour algorithm.The nonuniformly distributed data is given spefici boundary points or margin so that each data from the data set can be classified in proposed catergory and with the help of FCNN rule the dataset gets combined in various clsters.

Data structure and algorithm behind DOMIH method



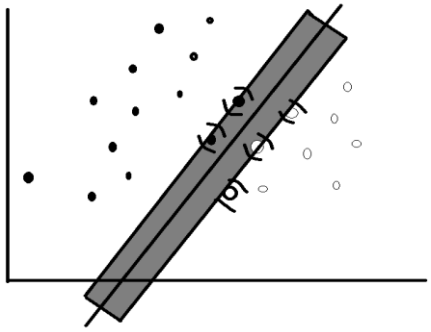
X. ALGORITHM / PROTOCOL / MATHEMATICAL INDUCTION / METHODS USED

1. *SVM(Support Vector Machine) method :*

For correct classification of mis-classified data,there are two options:

1) Get anew classifier

2) Use margin classifier i.e. classifier that has some margin



The classifer shown in figure is a margin classifier called as Support vector machine(SVM),

i) which avoids introduction of another classifier because of mis classification of some training data points near to decision boundary of binaryu binary classifier.

ii) which uses margin acting as classifier

Ideally this margin is expected to be maximum, so SVM is also called as maximum margin classfier Traing data points(input vectors)which are nearest to classifier are called as support vector and hence this maximum classfier is named as support vector machine. Decision boundaryof support vector machine is given as t=w.x

Where t=some there hold according to which input instances is classified to class {+1,-1}.

Mathematically, margin of SVM is given as $\frac{m}{||w||}$ which is (the smallest) space of support vector from decision boundary measured in the direction of w.

Note that W is the weight vector perpendicular to t=wx.

Note that for all feature vector X_i belonging to positive class,

$t_i = w.x_i$ +m

Similarly for all feature vector X_i belonging to negaitive class,

$t_i = w.x_i$ -m

Target of SVM is to maximise sum of positive margin and negative margin.This can be mathematically given as

Maximize $(m^+ + m)$

Maximize $(\frac{m}{||w||} \frac{m}{||w||})$

Maximize $\frac{2m}{||w||}$

2. *Triangle In equality:*

The algorithm is used to reduce the number of calculations space memory storage by decreasing training dataset which are not close to the test sample.

The algorithm can be described as follows:

Calculate the space between each training pixel to the other

• If there are n samples, this would mean space calculations.

For each test sample

i) Calculate the space from the first training sample as dn. This would be the current minimum space.

ii) Calculate the space from the second training sample (p) as dp.

iii) If dp < dn assign dn =dp

iv) For each remaining training sample(c) If space between the sample c and sample p measured as dcp meets

dp - dn < dcp < dp + dn

Calculate space from test sample to the sample c as

dp If dp < dn,

Assign: dn =dp

Else, skip this training sample.

Stop if there are no more training samples

3. FCNN Algorithm:

1. Go through the training set, removing each point in turn, and checking whether it is recognised as the correct class or not

I.If it is, then put it back in the set

II.If not, then it is an outlier, and should not be put back

2. Make a new database, and add a random point.

3. Pick any point from the original set, and see if it is recognized as the correct class based on the points in the new database, using kNN with k = 1

I.If it is, then it is an absorbed point, and can be left out of the new database

II. If not, then it should be removed from the original set, and added to the new database of prototypes

4. Proceed through the original set like this

5. Repeat steps 3 and 4 until no new prototypes are added

The algorithm FCNN rule (i) terminates in a finite time, (ii) computes a training set constant subset, and (iii) is order not dependent.

ACKNOWLEGDEMENT

This work is supported by Nutan Maharashtra Institute of Engineering and Technology, TalegaonDabhade, Pune.A high contribution of Guide Prof. Lomesh Ahire (Assistant Profressor at NMIET), Prof.PramodPatil (Project coordinator). Also guidance and Support from Principal Dr.RejendraKanphade prove beneficial.

CONCLUSION

We have introduced a data-oriented multi-index hashing method to solve the problems of the state-of-the-art methods efficiency losing in handling non-uniformly distributed codes and without ranking for accurate search. The observed superior learning speed of the new method is substantiated bythe learning behavior comparison. This work can be extended in several ways, e.g., studying the impact of different metrics on the FCNN rule, and the acts of FCNN-based hybrid process.

REFERENCES

(1) Fast Search in Hamming Space with Multi-Index Hashing Mohammad Norouzi Ali Punjani David J. FleetDepartment of Computer ScienceUniversity of Torontofnorouzi,alipunjani,fleetg@cs.toronto.edu

(2) Distributed Hash Tables in P2P Systems - A literary surveyTimo Tanner Helsinki University of Technologytstanner@cc.hut.fi

(3)Fast Condensed Nearest Neighbor Rule FabrizioAngiulliangiulli@icar.cnr.itICAR-CNR, Via Pietro Bucci 41C, 87036 Rende (CS), Italy

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com