Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp411-416)

Country Wise classification of Human Names

Raju Balakrishnan
India Software Lab, IBM™
Bangalore, India

Abstract:-Person names in a country follow a particular statistical trend and names of a large set of i country are derived from a set of names having smaller cardinality. The frequency distribution of pe different countries varies from each other. The intuitive ability of humans to guess the country of origin of his name is based on these facts. It is possible to design a data mining approach for deciding the country person from his name-using the first name and second name as the only independent parameters-and such a range of applications. But this is an unexplored problem, complexity and lack of information about human different countries may be the reason. In this paper we try to tackle this problem with two data mining algo we try a k-nearest neighbor classification for first names and second names, followed by a rule based de The algorithm is trained and tested on person names from nine countries. This method shows accuracy up to of ten countries. Secondly, we try an unsupervised method to improve the knowledge base of the system a algorithm can effectively handle the scenarios of 1) a small training set. 2) Apriori probabilities of w unknown at training time. The method shows accuracy up to 64% for nine countries.

Key Words:-Levenshtein distance, k-nearest neighbor classification, Etymology, Genealogy, Data Mining.

1. Introduction

Though person names are generated in a non-systematic and distributed process, which is influenced by individual preferences and emotions, the names in a population shows strong statistical trends. Ability of a human being to guess the country of origin of names in his region of residence with good accuracy is a clear outcome of this fact. A tool with this capability has wide range of applications in information retrieval, data mining, non-obvious relationship identification etc. An analytical study to find out the approximate percentage of PhDs issued to Chinese researchers in US or an e-commerce application to separate all Indians from an e-mail listing to send advertisements of a B2C portal for Indian food items are some examples of potential end user applications.

Though the applicability of this method is wide spread and obvious, the problem is an unexplored one. Reasons could be the complex nature and lack of systematic information available about the problem space. Unlike other domains such as biology, chemistry etc where name based classification has already been tried; the person-name space has evolved in a more random manner, with no scientific rigor, influenced by sociological and human factors.

Hence we need take into account less systematic and more heuristic factors while formulating a method of classification. One factor we noticed is, an average human adult can easily identify the region of origin, given a sample name from the geography she has for some time, than names originating in the world. This is essentially the problem with the names in the region. In other w stored a number of names of that particula memory. We hypothesize that a kno combined with excellent inexact patt capability of the human brain gives this abi This suggests an instance based learn followed by a pattern matching

Recent research in sociology thr this space. Bentley et al [4] concludes that t of first names in a population can be explai random copy of names by individuals from population and a random drift, and this pr power distribution of names in the popul and Marubia [5] show that family name populations' exhibit power law distribution values for exponents. Power law distribu names and family names imply that a small from a population can represent names of la individuals, and guide us to using a sample to tackle the problem. The second names a through the vertical genealogical lines, fro descendents, and less susceptible to change populations in countries are likely to cop from other populations also in today's con family names are more reliable in identify of origin of an individual as they are copie to progenies. This means that family nam

given more weightage compared to first names in deciding region of origin of a person name.

Considering the above factors, we devised a k-nearest neighbor classification algorithm for country wise classification of person names. Only two independent parameters, first and second names, are used. This keeps the dimensionality low, and helps to keep the training set size low. Another important factor noticed is the implicit information contained in association between first name and family names in a person's name. Association of first name to a particular country suggests that the second name also like to be associated to the same country and vise versa. This property is utilized for unsupervised expansion of knowledge base during run time.

2. Problem Definition and Analysis

In short, the problem is to identify the country of origin of a person name, i.e. to classify person names according to the country of origin. The problem is stated below in detail.

L = Learning set, which is a set of 2-tuples of names and countries of origin of those names, selected randomly from a list of names from N countries.

T = Test set of M names selected randomly from a disjoint set of persons from which L is drawn, each of them belonging to one of the N countries from which training set is chosen.

Problem 1 (Supervised): Train with the <name, country> 2-tuples from L, Identify names in test set as belonging to one of the above N countries , or belonging to none of them.

Problem 2 (Unsupervised): After an initial supervised training with an initial training set L', expand the knowledge base using T.

Output: 2-tuples <country, confidence > for each input name, where confidence increases with probability of correctness of classification.

Applying Bays criterion for maximum accuracy [21], any name to should be mapped to the class having the maximum probability for it to be a member. For other classes, the name will be classified erroneously. Note that the error can only be minimized and can not be eliminated irrespective of method of classification.

Scaling Bays's criterion to many classes, for maximum accuracy, x should be classified as belonging to C_j if

$$f_j(x) \geq \forall_i f_i(x) \qquad (1)$$

If the apriori probabilities of classes are not same this condition to be modified as

x should be classified to C_j if

$$P_j(C_j) f_j(x) \geq \forall_i P_i(C_i) f_i(x) \qquad (2)$$

Where $P_i(C_i)$ is apriori probability of C_i.

This can be achieved by adjusting the ratio of training samples from different classes proportional to the apriori probabilities of the corresponding classes in the test set. In our problem $P_i(C_i)$ is proportional to the ratios in which the person names belonging to different countries appear for classification in run time environment, which may not be predictable at training time.

3. Related Work

In sociology, distributions of first and second names in a population are found to be following power-law [3, 4]. The stochastic processes involving a random copying of names from members of the population along with a random drift is shown to be accurate model for the propagation of first names in a population [6]. A stochastic process involving vertical copying of family names from parents to children with eliminations of family names with mortality can explain the power law distribution in the frequency of family names [5]. The etymology sites list the common names in a particular country [3]. The US survey results are available sorted in the order of popularity of name [2]. US social security administration provides a list of popular baby names for a century [8]. Social Security Administration has published a brief study of distribution of given names of social security holders [7]. Classifying names into language classes using a hidden markov model based techniques were discussed by B T Oshika et al [20].

There is no prior works dealing with the problem of name in classification of person names to the best of my knowledge. T Grass et al describe a classification of proper names for compiling a multilingual database of names [9]. M Torii et al. explored different information sources helpful for classifying names of different biological entities [12]. M Collins and Y Singer describes two unsupervised algorithm for named entity classification-classifying named into classes like person, organization, and location etc-based on contextual and spelling information [13]. J Kazama et al explores using support vector machines for biological named entity classifications [16].

4. System and Method

4.1 Data Set

The list of person names was extracted randomly from the employee database of a multinational organization. The training set and test set were chosen from disjoint set of persons. The method was tested for G7 (United States, United Kingdom, France, Germany, Canada, Japan and Italy) countries and two other countries with largest population in the world-India and China. The availability of sample set was prime

considerations while choosing these countries. Together these nations cover more than half the population of the world. Equal number of names is chosen from each country for both test and training set. A test set size of 9000 (1000 from each country, total 9 countries) was used and classification performance is found out for different training set sizes and methods.

The middle names and initials are removed from names and 3-tuples of <first_name, second_name, country > are used for training set. Notice that the knowledge base consists of randomly chosen names. Carefully formed training set comprising of most popular names in their respective countries [2] is likely to further increase the accuracy of classification. But such a method will not be feasible for populations for which this information is not available. For the employee databases we used to extract our training and test set, pollution is high. For example many of the employees in US organizations may be Indians, which is not a true representative general of US population. No refining of the samples was performed to avoid pollution, as lack of information about the origin of person made any kind of refining impossible. The same number of person-names is used in training set from all nine countries, for both training and test sets to make the apriori probabilities of different classes same.

The required properties of a training set for optimum classification accuracy based on Bayes theorem of optimum accuracy in the context of the problem is explained below.

Let the test set be represented by the matrix

$$\text{Re}\,gions \rightarrow$$

$$Names \downarrow \begin{bmatrix} w_{11} & w_{12} & w_{13} & . & . & w_{1n} \\ w_{21} & w_{22} & w_{23} & . & . & w_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ w_{m1} & w_{m2} & w_{m3} & . & . & w_{mn} \end{bmatrix}$$

Where w_{ij} represents the number of occurrences of $Name_i$, belonging to $Region_j$, in the test set.

And training set be,

$$\text{Re}\,gions \rightarrow$$

$$Names \downarrow \begin{bmatrix} t_{11} & t_{12} & t_{13} & . & . & t_{1n} \\ t_{21} & t_{22} & t_{23} & . & . & t_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ t_{m1} & t_{m2} & t_{m3} & . & . & t_{mn} \end{bmatrix}$$

Where t_{ij} represents the number of occurrences of $Name_i$, belonging to $Region_j$, in the training set.

$Name_i$ is always mapped to one country. For optimum accuracy $Name_p$ should be mapped to $Region_q$ where

$$\forall i[w_{pq} \geq w_{pi}] \qquad (3)$$

This criteria is satisfied if,

$$\forall i \neq q[t_{pq} > t_{pi}] \qquad (4)$$

This condition is satisfied if the apriori probabilities of classes in the test set are same as that of the training set.

4.2 Training Methods

We tried two training approaches, namely supervised and unsupervised. For supervised training the training set and test set are compiled using 3-tuples of <first name, second name, country>. The first names and second names are added as separate unrelated instances in the knowledge base. 2-tuples of <name, country> are added in the database of the corresponding country. Here, a name can be either a first name or second name. We do not make any distinction between a first-name instance and a second-name instance in knowledge base. While adding a name for the first time, the count set to one. If the same name is hit again in the training set associated with same country, the count is incremented by one for the association between country and person.

Unsupervised training works based on the implicit information contained in the first name-second name association to build the knowledge base. A smaller initial knowledge base in created using supervised learning. When a first name is identified as associated to a particular country in the knowledge base, an association is established between the second name and the country in the knowledge base. Similarly if a second name found associated with a country and does not find the first name in the knowledge base, an association between the first name and the respective country is added in the knowledge base. If the system finds both first and second name in the knowledge base associated with the same country the count values of both the first and second names are incremented by one. Similarly, either first name or second name is associated with only one country; the count value of the association is increased by one step. Using this learning method the system will be able to expand its knowledge base and learn to classify better and can evolve into a better system with experience, like humans.

Unsupervised training has the advantage that the test set will always be a true representative of the training set as they are the same set of names. This is particularly important if apriori class probabilities of working set of the system are unknown at training time. According to bays criteria for minimum error discussed above, the apriori class probabilities for different countries need to

be same in training and test set for better classification accuracy. If the classifier is working in a dataset in which the frequency distribution of names against the countries is different from that of the supervised training set with which the classifier is initially trained, the system will be able to adapt over a period of time to the frequency distribution of the working set by unsupervised learning and to improve accuracy.

4.3 Classification Methods

Four steps involved in classification are,

1. Pattern Search in knowledge base for first name
2. Pattern Search in knowledge base for second name
3. Assigning confidences for each association.
4. Choosing an association of country and person names from results.

For pattern search we used k-nearest neighborhood approach. Levenshtein distance (Edit Distance) [17] is used as measure of distance between two names. Only names starting with same letter as the starting letter of the given name is compared for nearness, since the starting letter is very important in deciding how the name sounds as suggested by phonetic coding algorithms like soundex [18] and NYSIIS [19]. Note that an edit distance of zero means an exact match and there can be only one name associated with one country in knowledge base satisfying this condition. For distances one and above (distance zero is an exact match) there can be multiple names in knowledge base equidistant from a given name. Confidence for an association is calculated as:

$$confidence_i = \frac{count_i}{\sum_{j=1}^{N} count_j} \qquad (5)$$

Where N is total number of associations found for that name and $count_i$ is number of times that particular name is found associated with the country in training set, which we keep track of as mentioned in section above.

For a <first name, second name> pair, the database is searched for both first name and second name. If there is only one association in knowledge base-for first name or second name-person name is associated with that country. If there is more that one association for only one name-first name or second name-the country having maximum confidence value is chosen as associated country. If there is one or more association for both first name and second name, we find out the intersection of sets of countries associated with first name and second name. If there is only one country in the resulting set that name is associated with that country. If there is more than one country in the intersection, country for which maximum value for sum of confidences (sum of confidence for first name and second name association) is chosen. If there are zero members in intersection, we check if there is an entry in set of associations for first name with confidence value of greater than two times of maximum confidence in set of associations for second name (choice of value two is empirical) If the condition is satisfied, association of first name with maximum confidence value is chosen. Otherwise association of second name with maximum confidence is chosen.

4.4 Implementation Details

The program was implemented in java™ programming language. For searches with edit distance zero in training set a hash table based search is used. This is important in performance point of view. For large knowledge bases exact match is found for most of the names. Expected time for hash table search is O(N), assuming a hash table size greater than the number of entries and hashing function giving uniform distribution of entries, where N is the number of countries for which the classification is performed [14][15].

For comparisons having edit distance greater than zero, distance of all names starting with the same letter is determined. The time complexity of this method is O(NM) where M is number of names for each country in knowledge base, which is obviously greater than that of exact string matching described above.

5 Results and Discussion

We tried prediction on a test set size of 9000(See the section on data set for details of test set). The following 3 figures represent results for

1. Supervised Training (Fig.2)
2. Unsupervised Training (Fig.3)
3. Effect of varying initial supervised learning set size on results of unsupervised learning (Fig.4).

In each figure two graphs are plotted, one graph showing results of nearest neighborhood method, and second graph showing results in which the pattern matching is solely exact string matching method. If there is no exact match found for the name in knowledge base, the name is classified as unclassified. Note that this will give better precision.

$$\text{Accuracy} = \frac{\text{No : of names mapped to the country person belongs}}{\text{Total no : of person names in test set}}$$

Note: The X axis represents size of the training set from one country; total training set size will be number of countries multiplied by this value.

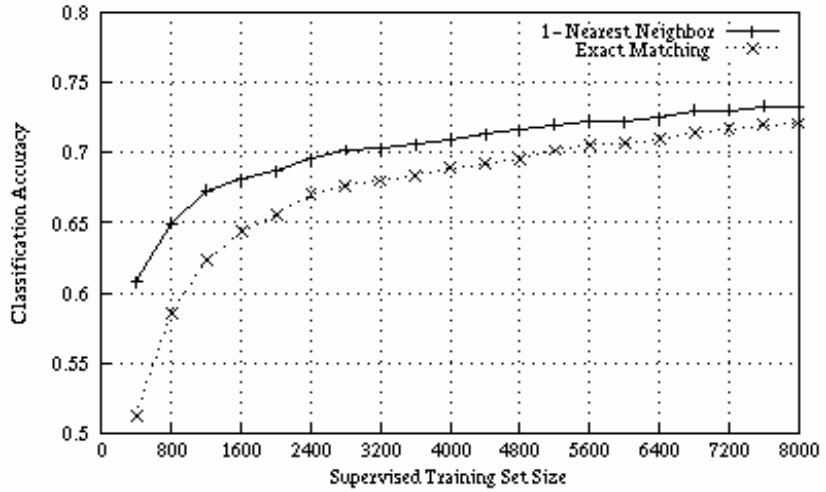

Fig.2: Supervised learning results show a monotonic increase in accuracy with training set size. The accuracy for exact matching approaches that of 1-nearest neighborhood for large data sets. The graph suggests that if large data sets are available the exact pattern matching is an excellent candidate, especially for performance critical systems (See performance discussion in implementation section) requiring high precision

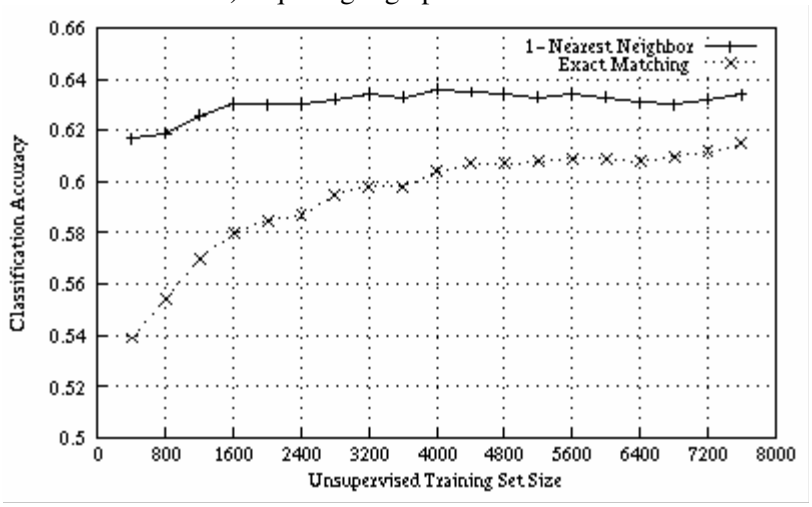

Fig.3: Unsupervised training results for nearest neighborhood string matching shows only small increase in accuracy with training set for 1-nearest neighborhood. The accuracy is optimal around a training set size of 4000 and slightly decreases after that. This is due to increase in misclassified samples. But he exact string matching shows a monotonic increase in accuracy with training set size and approaches performance of 1-NN method for large data set.

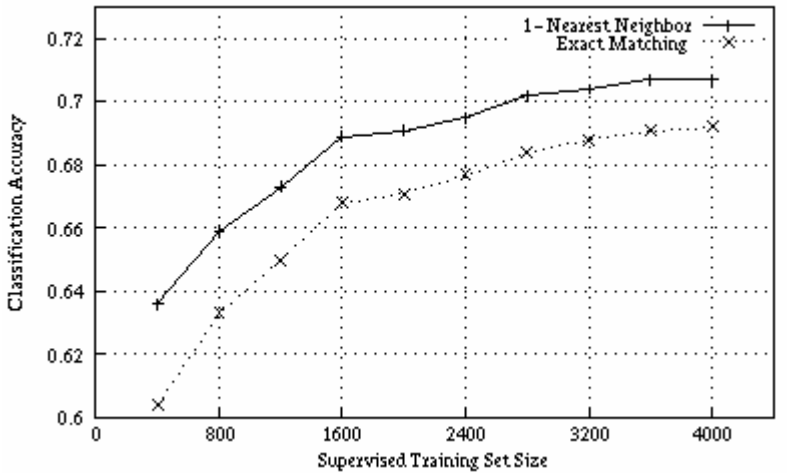

Fig.4: For this graph a fixed unsupervised training set size of 4000 (for each country) is used and initial supervised training set size in varied form 400 to 4000. The accuracy shows a monotonic increase.

6 Summary and future work

In a test set selected from nine countries (These countries contains, 3.062 Billion, more than half of the world's population) method archives up to 73% accuracy in prediction. This shows applicability of the method for a set of counties of this size. The names can be used as one of independent parameters in identifying a person's origin, as well as in cases where no other more authentic information is available about the person's origin. Incorporating information from cytology and genealogical about name distribution in countries may improve the result. Combination of instance based non-parametric method described in this paper and phoneme based methods for language wise classification of names may give better results with knowledge of language prevalent in each country.

References

[1] List of countries by population. http://en.wikipedia.org/wiki/List_of_countries_by_population

[2] U.S Census Bureau 1990 names list, http://www.census.gov/genealogy/names

[3] Etymological site listing the popular names in each country. http://www.behindthename.com

[4] M W Hahn, R A Bently, *Drift as a mechanism for cultural change: an example from baby names*, Proceddings of Royal Society London, Biology Letters Volume 270, 2003, pp.120-123.

[5] Zanette,H Damian, Manrubia, *Vertical transmission of culture and the distribution of family names*, Physica A 295, 2001 1-8.

[6]R A Bentley, M W Hahn, S J. Shennan, *Random Drift and Cultural Change*, In Proceedings of Biological Sciences, Royal Society London, Volume 271, 2004, pp.1443-1450

[7] Name Distribution in social security area, US Social Security Administration11, http://www.ssa.gov/OACT/NOTES/note139/original_note139.html .

[8] US Social Security Administration, Popular Baby Names, http://www.ssa.gov/OACT/babynames/

[9] T Grass, D Maurel, O Pinton, *Description of a multilingual database of proper name,* Proceedings of the Third International Conference on Advances in Natural Language Processing,2002, pp.137-140.

[10]S Sekine, R Grishman et al. *A Decision Tree Method for Finding and Classifying Name in Japanese text,* In Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

[11] S Wright, *Evolution in Mendelian Populations,* Genetics, Volume 16, 1931,pp.97-126.

[12] M Torii, S Kamboj, K Vijay-Shanker, *An Investigation of Various Information Sources for Classifying Biological Names,* 41st Annual Meeting of the Association for Computational Linguistics, July, 2003.

[13] M Collins, Y Singer, *Unsupervised Models for Named Entity Classification*. In Proceedings of EMNLP 1999.

[14] M A Weiss, *Data Structures and algorithm analysis in C*, Addison-Wesley

[15] Aho, Hopcraft, Ullman ,*The Design and Analysis of Computer Algorithms*, Pearson Education Asia.

[16] J Kazama, T Makino, Y Ota, J Tsujii. *Tuning Support Vector Machines for Biomedical Named Entity Recognition*. In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, 2002.

[17] P E Black, *Algorithms and Theory of Computation Handbook,* CRC Press LLC, 1999, "Levenshtein distance", from Dictionary of Algorithms and Data Structures, ed.

[18] Soundex Algorithm http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm

[19] P E Black, "NYSIIS", from Dictionary of Algorithms and Data Structures, P E Black, ed. http://www.nist.gov/dads/HTML/nysiis.html

[20] B T Oshika, B Evans, F Machi, J Tom. *Computational Techniques for improved name search*. In Proceedings of Second Conference on Applied Natural Language Processing, 1998, pp 203-210.

[21] Prof. Nitin Patel. MIT Opern Courseware, 15.062, Data Mining Spring 2003.

Note

[1]IBM is a registered trademark of IBM Corporation in the United States, other countries, or both.

[2]java is a trademark of Sun Microsystems in the United States, other countries, or both