Centroid based Categorization Approach for Extraction of Body Sensor Network Data

Setu Ku. Chaturvedi
setu16@gmail.com
TCT, Bhopal

Basant Tiwari
basant_tiwari@rediffmail.com
TCT, Bhopal

Abstract: Monitoring human activities using wearable wireless sensor nodes has the potential to enable many useful applications for everyday situations. The long-term lifestyle categorization can greatly improve healthcare by gathering information about quality of life; aiding the diagnosis and tracking of certain diseases. The deployment of an automatic and computationally-efficient algorithm reduces the complexities involved in the detection and recognition of human activities in a distributed on Body sensor network server. Directory service is a useful aid human looking for information on Network Data. A directory services is a pre-categorized list of topics containing many links for each topic. However, most directory services are maintained manually now and face many drawbacks. Therefore the task of automatic categorization of new data into the topics of directory services becomes very necessary. BSN data categorization is more difficult due to a large variation of noisy information embedded in Sensor network data. This paper suggests a new Centroid based approach for Categorization for BSN data. We further introduce a new algorithm through centroid based approach for extraction of BSN data categorization and show that it achieves about more improvement over other BSN data categorization methods. Experimental results show that our proposed Centroid-based BSN data categorization algorithm achieves an approximately 13.8% improvement for BSN data categorization algorithm.

Keywords: Centroid-based categorization, BSN data categorization, Centroid.

I. INTRODUCTION

The development of small sensor platforms from relatively inexpensive, commercially available components has created fascinating new opportunities for data collection. The sensor nodes can communicate wirelessly, have limited storage capabilities, and are often deployed in networks. These sensor networks have been used for a wide variety sensing tasks from individual health monitoring to large-scale environmental Sensing. The sensors measure environmental variables including temperature, humidity, force, acceleration, and heartbeat. In most applications, a light-weight embedded sensor node is expected to acquire physical measurements, perform local processing and storage, and communicate over a short distance. Sensor networks that understand human actions are expected to be useful for numerous aspects of everyday life. Human motion recognition using wireless sensor networks can employ data from either environmental or on-body sensing devices [1].

BSN Server has a huge source with a lot of information valuable for human. However, people need aid to retrieve this information due to the vast amount and spread distribution of BSN data since it is collected from various patients' bodies. The common kinds of services available on BSN data is to extract patient's data from database contains all the data about all the patients wear the BSN.

Proposed work introduces the directory services for categorization of BSN data. A directory service is a pre-categorized list of topics (subjects) Containing many links for each topic. These links may direct to a specific data but Most of them link to lots of data. We cite this definition from Nassau Library System [3]: A BSN data is a group of heterogeneous data collected from various patients bodies wear BSN and stored and managed by a single server. BSN data can range in size from as few as one data to a vast number of data depends on the number of sensor attached within the BSN. Thus, the information presented on a BSN data might belong to various categories but follows a common purpose.

BSN data directory and most other commercial directory services are still set up manually. We will examine a symbolic directory service; first, the BSN data must choose an appropriate category or suggest a new one. Then he or she needs to submit the BSN data to directory. First, it will be very time consuming to extend a large directory service. Second, the links for category will be soon out of date. Third, a manual directory service will be hard to satisfy some specific patient data. Fourth, the consistency of categorization is hard to maintain since different human experiences are involved. Clearly, we can overcome these drawbacks by using

an automatically classifier. Although other data classifiers are well researched for data, categorization of the whole BSN data is still being investigated. In this paper, we introduce a Centroid based scheme that will improve the BSN data categorization task. We select BSN data for Centroid. Each BSN data will be represent by a set of feature vectors (information) and each category will be represent by a centroid-based feature vector. The new BSN data will be classified based on how closely its features matches the features of the current BSN data belonged to different category.

The organization of this paper is as follows. Section 2 contains a summary of related work in categorization of BSN data. Section 3 suggests our Centroid - based scheme. Section 4 describes centroid based BSN data categorization. Section 5 shows some experimental results. Finally, section 6 provides directions for future research.

II. RELATED WORK

The various categorization algorithms are well research for many years and fall under three general categories. The first category contains algorithms using retrieval techniques such as prototype-based classifier (Rocchio), k-nearest neighbor (KNN), centroid-based classifier [4][5]. The second category includes discriminative classifiers originally from machine learning such as decision-tree, support vector machines (SVM). The final category contains generative classifiers such as naïve Bayes classifier. A comparison between these methods is available in [6] considers class labels and the text of neighboring (linked) data[9] applies on BSN data though Centroid based categorization[12]. It introduces a new tree structure for the hierarchical BSN documents categorization. However, these methods focus on classify whole BSN data. Some approaches of classifying BSN data are introduced in [10]. They based on different representations of BSN data. First approach, a BSN data is represented as a single BSN data consisting of the union of all its data. Second approach, a BSN data is represented by a vector of topic frequencies. Final approach, a BSN data is represented by a tree of data with categories. In an approach using local text representation was stated. In a BSN, data is represented by a set of feature vectors using term frequency and the authors use a centroid-based categorization approach that has satisfactory results. However, this potential approach can be improved by our scheme, which is based on some Centroid approach for data.

III. CENTROID TECHNIQUE

Centroid is a major problem related to BSN data categorization. Generally the high dimensionality of the term space can make the classifier run slowly i.e. in this set of experiments we consider following approaches to the reduction of the dimensionality for the context of BSN data categorization.

- *Item frequency*, that consists in removing terms that occur less than times in the training set.
- *Topic frequency*, that consists in removing terms that occur in less than examples of the training set.

The simplest approach summarizes all data within a BSN data into one single feature vector and afterwards classifies this vector. However, this approach performed very poorly and thus, classifying a BSN data in the same way as a single data is an inadequate approach to BSN data categorization.

The approach offering the best results is named *topic frequency vector* approach (TFV-approach). To employ the TVF-approach for BSN data categorization, several preprocessing steps are needed. First of all, a set of training BSN data has to be acquired employing a directory service that is easy, since there are already leaves in the topic category, mostly linking to relevant sites. To attain the data representing these BSN data, one could sensed and stored all data at server, but it is usually enough to restrict the number of data to maximum number. During categorization an incremental classifier as described below can achieve this restriction.

The next step is to find a proper set of data classes for each BSN data. The set of data classes contain data that describe data being typical to occur in a BSN data of a certain class. For example, ‘Body temperature” might be a meaningful data class containing temperature of all the patients. To distinguish all BSN data belonging to some other class, we employ a global other BSN data class that is not specified any further. One “other” data class describes this “other” BSN data class only. After definition, we have to generate training examples for each of the data classes. Therefore, we examine the data within the training BSN data and manually label the data with the proper data class. Thus, we receive the training set for the data classes, additionally to the needed set of training BSN data. Both of these steps tend to be very time consuming, since there is no approved general way for automatic generation of classes and training examples. For example,

determining BSN data classes and labeling enough training data took about 3 days for the data used in the evaluation of [4]. After generating the training set for the data classes, we train a naive Bayes classifier that is capable to label new unknown BSN data with the most likely data class. To classify a BSN data, we derive the so-called Topic Frequency Vectors (TFVs).

The vector space of topic frequencies is spanned by the set of BSN data classes. The idea is to count the occurrences of each data class from different BSN data. Thus, a TFV gives a brief overview of the BSN data classes that occur within a site. For categorization, we employ a second naive Bayes classifier trained on the TFVs derived from the training BSN data.

Categorization of an unknown BSN data is achieved in two steps[13]. First, we classify the data of the BSN data and therefore, derive the TVF. Afterwards we employ the second classifier to predict a Topic class from this TFV. In additionally described a variant for incremental categorization that does not employ all data of a BSN data.

The idea is to read a data from the BSN data given by the domain (Class) name only, follow the links and measure the quality of each path leading away from it. If this quality is not"relevant" enough, we prune the path. After all paths are pruned, the treated portion of the BSN data is usually a good representation for the complete data. In this incremental variant, the prediction of the second classifier is calculated during the traversal of the BSN data and is used to decide the quality of each path to be pruned.

IV. CENTROID-BASED CATEGORIZATION OF BSN DATA

The centroid of BSN data is a useful representative of a complete class in terms of KNN categorization [16]. We take up this idea and apply it to BSN data categorization. The idea of the centroid-based categorization is that each topic (BSN data *class*1) contains several groups of data that are some how related and can be extracted by a common representative, a centroid vector.

Generally, given some groups of related elements, a mean vector for the elements of each group is calculated. We call it centroid vector. A class containing some groups of related elements is represented by the centroid set as the set of all such centroid vectors. We cite the equation from [11] with a little notation change. Let *S* be a set of groups of elements e_i with vectors

$$v_{j,g_i}. \text{ Let } \pi_g(e_i) = \{v \mid f(v) = g \ \forall v \in e_i\}$$

be the restriction of e_i to group *g* where f is a mapping from a vector *v* to a group $g \in G$, the set of all groups. Then the centroid set *CS* of *S* is defined as:

$$CS(S) = \left[c_j \mid \forall j \in G, c_j = \frac{1}{\left|\bigcup_{\forall i} \pi_g(e_i)\right|} \cdot \sum_{x \in \bigcup_{\forall i} \pi_g(e_i)} x \right]$$

When applying this for BSN data categorization, we represent the content of a single data *p* by a feature vector and so a whole BSN data *W* by a set of feature vectors. To get an illustrative view, please refer to Figure 1 showing the idea to represent a sample BSN data class with a centroid set.

We now have two remaining problems. First, we need to choose an appropriate distance measure function. Second, we have to group the similar data within a BSN data class, a clustering task. We use a simple cosine function to measure the similarity between a data and a centroid vector. However, in the context of centroid-based BSN data categorization, Half Sum of Minimum Distances (HSMD) seems to be the most adequate distance measure between test BSN data and the centroid sets. For the detail explanation, please refer to [11]. Let *W* be a BSN data, let *C be* a centroid set and let $f : d\ P\ N \rightarrow$ be a mapping from P, set of all data and centroid sets, that returns the feature vector of $p \in P$. The *d(x,y),* the classic Manhattan function, is used to measure the similarity between two feature vectors due to their much simpler mathematical operations .

$$d(x,y) = \sum_{i=1}^{n} \left| \frac{x_i}{N_x} - \frac{y_i}{N_y} \right|$$

$$HSMD(W,C) = \frac{\sum_{w_i \in W} \min_{c_j \in C} d(f(w_i), f(c_j))}{|W|}$$

Clustering is well studied for many years with two main approaches, similarity-based and model-based. However, we have to consider some requirements when choosing a clustering algorithm. First, we don't know exactly the number of BSN data topics, so we can eliminate clustering approaches that require inputting the number of clusters (such as K Means).

Second, the chosen clustering algorithm must deal with noise since, in many cases, several data uncommon for the topic of BSN data they belong to. Third, we don't have any pre training set of BSN data. Fourth, a major issue is that directory service requires high rate of update in a dynamic environment so the clustering method must adapt it. Considering all above requirements, we choose *GDBSCAN* (Generalized Density-Based Spatial Clustering of Applications with Noise) to group training data within each BSN data class. For more detail about GDBSCAN algorithm please refer to [13].

So we summarize the algorithm to determine the centroid set for a BSN data class *Ci :*

1. Collect all data (their feature vectors) of the test BSN data of class *Ci* into one super set.
2. Determine clusters using GDBSCAN based on the set of feature vectors.
3. For each cluster, calculate the centroid vector and insert it into the centroid set of class *Ci*.

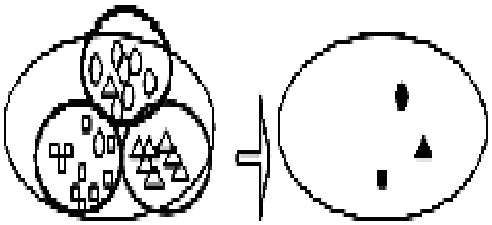

Figure :1 Centroid of a sample class

V. EXPERIMENTAL EVALUATION

At this time, there is not a standard data set of data for BSN data categorization tasks. We just selected 6 categories to avoid a too larger data set. For each category we randomly choose from 10 to 50 example of BSN data set. We also choose 50 BSN patient data from the other topics to make noise. Building a BSN crawler with breadth first traversal described in Section 3, we retrieve from 100 to 120 for each BSN data. Finally, my data set has 127 BSN dataset containing about 14,000 data.

We perform two experiments. First experiment is a binary categorization for each topic. The second is a 6-class categorization for all categories. The judgment based on precision and recall measurement [16]. We compare the naïve centroid-based categorization (Cent.) with the centroid based categorization using our *weighted term frequency* scheme (Cent.WTF).

For first experiment, there is not big improvement when using Cent.WTF. The results are displayed in Table 1.

However, in news topic, Cent.WTF increases the precision considerably. We paid attention that the Cent. performs weakly in these categories.

Category	Centroid		Centroid WTF	
	pre.	rec.	pre.	rec.
BP	.75	.82	.75	.82
Heart Rate	.85	.71	.85	.72
Body_Temp	.65	.88	.72	.85
Blood Oxigen	.86	.76	.87	.79
Hb	.77	.82	.75	.84
EMG	.79	.73	.76	.75

.

Table1. Comparison of Cent amd Cent WTF in binary classification with precision recall measurement.

For second experiment, to have a general view of the 6-class categorization performance, we just calculate the overall accuracy, which is the percentage of correctly classified instances with respect to all tested instances. As the results in Table 2, the Cent.WTF gives a better categorization.

Category	Centroid	Centroid WTF
	accuracy	accuracy
6 class	.71	.75

Table2. Comparison of Cent and Cent WTF in 6-Class classification with accuracy measumrement.

To sum up, the centroid-based classifier using our Centroid based selection scheme outperforms the naïve centroid-based classifier.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we suggest a centroid feature selection scheme to improve the performance of the centroid-based BSN data categorization. This will help in automatically maintaining commercial BSN data directory services [14]. The experimental evaluation shows that this is a potential approach. In our future work, we will focus in three tasks. First, it is clear that not all features contributed equally in distinguishing topics. So we need a scheme to adjust the feature weight. Moreover, when some BSN data was inserted into the topics, the centroid will change

as a result and may be the feature weight need to re-adjust. So we will focus to build an iterative feature weight algorithm. Second, the centroid-based BSN data categorization now is term-based approach. However, the semantic links of data belonged to a BSN data also have an important meaning. We will propose a hybrid approach that uses both term-based and Co-term based methods when performing clustering groups of related data of a BSN data.

REFERENCES

[1] K. Venkatasubramanian, G. Deng, T. Mukherjee, J. Quintero, V Annamalai and S. K. S. Gupta, "Ayushman: A Wireless Sensor Network Based Health Monitoring Infrastructure and Testbed", In Proc. of IEEE DCOSS June 2005

[2] S.-D. Bao, L.-F. Shen, and Y.-T. Zhang, “A novel key distribution of body area networks for telemedicine,” in Proceedings of IEEE the International Workshop on Biomedical Circuits and Systems, pp. 1–20, Singapore, December 2004.

[3] Nassau Library System. (http://www.nassaulibrary.org)

[4] E.H.Han, G.Karypis, “Centroid-based Document Categorization: Analysis and Experimental Results”, *Proc. 4th PKDD 00*, Lyon, France, 2000.

[5] S.Shankar, G.Karypis, “Weight adjustment schemes for a centroid-based classifier”, *Text Mining Workshop KDD*, 2000.

[6] Y.Yang, X.Liu, “A re-examination of text categorization methods”, *Proc. of the 22nd annual international ACM SIGIR*, 1999.

[7] S.Chakrabarti, B.Dom, P.Indyk, “Enhanced hypertext categorization using hyperlinks”, *Proc. 17th ACM SIMOD*, New York, US, 1998.

[8] M.Craven, D.DiPasquo, D.Freitag, A.McCallum, T.Mitchell, K.Nigram, S.Slattery, “Learning to Construct Knowledge Bases from the World Wide BSN”, *Artificial Intelligence*, Elsevier, 1999.

[9] D.Shen, Z.Chen, Q.Yang, H-J Zeng, B.Zhang, Y. Lu, W-Y.Ma, “BSNdata Categorization through Centroid”, *Proc. ACM SIGIR’04*, Sheffield, UK, 2004.

[10] M. Ester, H-P. Kriegel, M Schubert, “BSN data Mining: A new way to spot Competitors, Customers, and Suppliers in the World Wide BSN”, *Proc. 8th ACM SIGKDD 02*, Alberta, CA, 2002.

[11] H-P. Kriegel, M Schubert, “Categorization of BSN data as Sets of Feature Vectors”, *Proc. of the IASTED International Conference*, Austria, Feb. 2004.

[12] W.Wong, A.W.Fu, “Incremental Document Clustering for BSN Data Categorization”, *Proc. of 2000 International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000)*, Japan, 2000.

[13] J.Sander, M.Ester, H.P.Kriegle, X.Xu “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications”, *Data Mining and Knowledge Discovery*, 1998.

[14] J.M.Pierre, “On the Automated Categorization of BSN Sites”, *Linkoping University Electronic Press*, Sweden, 2001.

[15] D.Riboni, “Feature Selection for BSN Data Categorization”, *EURASIAICT 2002 Proceedings of the Workshops*, 2002.

[16] C.J.V.Rijsbergen, “Information Retrieval”, *2nd Edition*, 1979.