Оценка информативности алгоритмов ЦОС

Лабораторная работа по курсу «Системы цифровой обработки сигналов»

Оглавление

[Цель лабораторной работы 2](#_Toc163850736)

[1. Теоретические сведения 2](#_Toc163850737)

[1.1. Формализация задачи классификации сигналов 2](#_Toc163850738)

[1.2. Формализация задачи отбора информативных признаков 3](#_Toc163850739)

[2. Система анализа информативности 4](#_Toc163850740)

[3. Задание 6](#_Toc163850741)

[4. Методика выполнения 6](#_Toc163850742)

[5. Контрольные вопросы 7](#_Toc163850743)

[6. Литература 7](#_Toc163850744)

[Приложение 1. Фрагменты данных и результатов в системе ICalc.mcd 8](#_Toc163850745)

# Цель лабораторной работы

При проектировании систем распознавания сигналов и систем диагностики на основе обработки диагностических сигналов разработчики часто выбирают либо тривиальные алгоритмы, либо сложные алгоритмы, оказавшиеся эффективными в других приложениях, либо просто привычные или удобные для программирования алгоритмы. Цель лабораторной работы – обосновать необходимость объективного подхода к выбору алгоритма цифровой обработки сигналов для решения задачи классификации сигналов и синтеза образов сигналов.

# Теоретические сведения

## Формализация задачи классификации сигналов

Сигналы регистрируются на интервале с известной частотой дискретизации , соответственно, шагом , индекс и моменты отсчетов сигнала – , .

Источник может находиться в одном из *M* состояний , в задаче диагностики часто рассматривают только два состояния – исправное (**C**) и неисправное (**B**). Требуется по наблюдению сигнала отнести сигнал к одному из *M* классов состояния источника, т. е. получить оценку состояния . Каждое состояние представлено подмножеством сигналов обучающей выборки . Подмножество сигналов, наблюдаемых в состоянии источника , образует класс сигналов . Таким образом, каждое состояние источника связано с классом сигналов . Мы надеемся, что сигналы класса достаточно похожи между собой и достаточно отличаются от сигналов класса , чтобы можно было использовать один из множества алгоритмов классификации. Решая задачу классификации сигналов, т. е. отнесения их к одному из классов , мы получаем оценку состояния источника .

Сигналы обрабатывают, концентрируя информацию о состоянии источника в векторе признаков **F** (**образе**):

**,** (1)

где n – число признаков, – функционал, значения которого для сигналов одного класса должны лежать в одной области, а для сигналов разных классов в разных областях, желательно максимально удаленных одна от другой. Функционалы (алгоритмы обработки) подбираются так, чтобы выполнялось условие «хорошей» кластеризации **образов** сигналов, т. е. для всех **классов** сигналов выполнялось условие:

(2)

– среднее расстояние между сигналами одного класса;

– среднее расстояние между сигналами разных классов.

Таких функционалов может быть определено несколько, , они могут принадлежать как одному семейству: спектральные плотности мощности на разных частотах [1], вероятности некоторых событий, частоты характерных последовательностей в сигналах, так и семействам разного типа [2].

Найденные преобразования сигнала (алгоритмы обработки) используются в рабочем режиме для **оценки** **состояния** или класса:

(3)

Назовем величину отношения количества отсчетов сигнала к размерности признакового пространства

(4)

*коэффициентом редукции сигнала*. Редуцирование сигнала к вектору признаков не только упрощает решающее правило, но и повышает информативность представления (**образа**) состояния источника.

**Образы –** векторы признаков, характеризующих объект.

Задача оценки информативности диагностических признаков является важным этапом разработки любой системы распознавания или автоматического анализа сигналов. Часто в качестве критерия информативности признаков выбирают вероятности правильного распознавания и ошибок первого и второго рода, оцениваемые по испытаниям классификатора в выбранном признаковом пространстве, а также системы распознавания в целом. При неудачном выборе признаков синтез такой системы оказывается напрасной работой. Таким образом, оценка информативности признаков и объективный выбор алгоритмов вычисления признаков до начала разработки системы распознавания или обработки сигналов имеет очень большое практическое значение.

Решение задачи оценки информативности признаков и признаковых пространств осложняется отсутствием единого критерия информативности. Решающим критерием является, разумеется, величина потерь. Но для ее оценки требуется спроектировать систему обработки сигналов и/или систему распознавания, а это как раз то, чего следует избегать, пока нет оценки информативности выбранных признаков. Известны критерии «трудности распознавания» - среднее значение евклидова расстояния между математическими ожиданиями всех пар сочетаний образов; разность между двумя ожидаемыми величинами логарифмов отношения правдоподобия векторов образов (дивергенция), энтропия распределений плотности вероятности образов в многомерном выборочном пространстве. Для семейства однотипных признаков полезен критерий - отношения среднего межклассового расстояния к среднему внутриклассовому расстоянию. Именно он выбран за основу в лабораторной работе.

## Формализация задачи отбора информативных признаков

Система отбора информативных признаков в задаче распознавания состояния сложного источника реализует следующий алгоритм.

Исходные данные представлены в виде таблицы **Th**, строки которой соответствуют образам , а столбцы – признакам . К таблице добавляется еще один столбец, его элементами являются индексы классов , к которым относятся образы обучающей выборки . – номер класса, к которому относится образ .

Итак, *K –* количество образов  в обучающей выборке, а *n* – количество признаков, размерность пространства образов. Индекс  - номер признака или координаты пространства образов. Элементами таблицы исходных данных являются значения признаков  .

**,** (5)

где  **–** характеристический вектор обучающей последовательности образов :

, (6)

и где  – матрица образов.

Рассчитывается матрица  расстояний между образами, на основе этой матрицы можно оценить потенциальные ошибки распознавания по методу ближайшего соседа (МБС) – случай, когда минимальное расстояние от рассматриваемого образа класса m относится к образу, не принадлежащему классу m.

Расстояние или функция расстояния понимается как отображение , удовлетворяющее условиям:

1. 
2. **.**

Для индексации функций пар образов используются индексы 

Далее рассчитываются внутриклассовые расстояния и расстояния между классами , которые могут служить критериями информативности, поскольку характеризуют компактность расположения образов одного класса и удаленность образов разных классов в пространстве образов (признаковом пространстве).

Наконец, рассчитывается основной критерий информативности признакового пространства **F** заданной размерности как отношение среднего межклассового расстояния к среднему внутриклассовому:

 (7)

Информативные признаки и пространства отбираются по результатам вычисления информативности пространств образов, получаемых прямым перебором.

Остановимся на случае, когда образом объекта является непосредственно сигнал. Матрица **X** теперь содержит отсчеты сигналов, N – количество отсчетов:

. (8)

Если признаковое пространство обычно имеет размер 2 – 20, то длина выборки сигнала составляет несколько десятков тысяч. Отбор признаков в такой постановке нецелесообразен, но общая оценка информативности сигналов и методов их обработки может оказаться полезной.

Например, образом может быть спектральная плотность мощности, полученная с помощью преобразования Фурье по выборке сигнала длиной N на частотах, кратных коинтервалу Найквиста ω1. Тогда матрица образов равна

. (9)

На основе преобразования Фурье может быть построен образ фаз колебаний с частотами, кратными коинтервалу Найквиста

. (10)

# Система анализа информативности

Метод основан на понятии хорошего размещения образов в признаковом пространстве, введенном в теории распознавания. Мы говорим, что образы хорошо размещены, если образы, характеризующие объекты одного класса расположены компактно, расстояния между ними малы. Образы, характеризующие объекты разных классов, должны быть удалены друг от друга, расстояния между ними должны быть велики. Естественно вводится относительный критерий – отношение межклассового расстояния к внутриклассовому.

Система анализа информативности включает в себя алгоритм расчета внутриклассовых (внутримножественных) расстояний, алгоритм расчета расстояний между классами (множествами), алгоритм вычисления критерия информативности. Параметром системы является указание на способ вычисления расстояния: евклидово, Чебышева и др. Входными данными является таблица **Th**  (5) – таблица «класс – образ», выходными – значения внутримножественных (внутриклассовых) расстояний, расстояний между классами и значения критерия информативности. Подсистема отбора информативных признаков формирует списки признаков и признаковых пространств с заданными пороговыми значениями (mmx) критерия информативности.

Для расчета информативности используются следующие способы вычисления расстояний. Обозначим евклидово расстояние между точками признакового пространства, которые соответствуют образам  и  классов  и как . Евклидово расстояние в многомерном пространстве равно

, (11)

Здесь *i* - номер признака, n - количество признаков, размерность пространства образов.

Внутриклассовое Евклидово расстояние

,  (12)

где  – количество образов в классе .

Евклидово расстояние между классами равно:

 (13)

Отно­ше­ние среднего межклассового Евклидова расстояния в признаковом пространстве к среднему внутриклассовому расстоянию в признаковом пространстве **F** выбрано критерием информативности:

, (14)

Особенности представляют собой случаи двух и нескольких классов. Так, образы первого и второго класса могут быть отделены от третьего, а между собой пересекаться. В этом случае целесообразно использовать минимальное отношение межклассового Евклидового расстояния к сумме их внутриклассового расстояния:

. (15)

Для оценки информативности использовалась «Система оценки информативности диагностических признаков и признаковых пространств» [3]. Система оценки информативности диагностических признаков и признаковых пространств позволяет объективно оценить информативность каждого признака. Добавление признака с большой информативностью часто не увеличивает, а даже снижает общую информативность по критерию отношения межклассового расстояния к среднему внутриклассовому, поскольку увеличиваются внутриклассовые расстояния, а межклассовые расстояния меняются незначительно.

Система позволяет:

* подобрать набор информативных признаков для разных типов алгоритмов, как метрических, так и логических;
* выявить причины неинформативности признаков и признаковых пространств;
* определить, необходима или нет обработка сигнала, выбрать метод обработки;
* определить, необходимо ли вычислять характеристики сигналов;

Входные данные системы – матрица «класс –образ», выходные – информативности наборов признаков. Система анализа информативности включает в себя алгоритм расчета внутриклассовых (внутримножественных) расстояний, алгоритм расчета расстояний между классами (множествами), алгоритм вычисления критерия информативности. Параметром системы является указание на способ вычисления расстояния: евклидово, Чебышева и др., выходные данные – значения внутримножественных (внутриклассовых) расстояний, расстояния между классами и значения критерия информативности. Подсистема отбора информативных признаков формирует списки признаков и признаковых пространств с заданными пороговыми значениями критерия информативности.

Критерий информативности (или просто информативность) в приведенных ниже примерах – это отно­ше­ние среднего межклассового Евклидова расстояния к среднему внутриклассовому расстоянию в признаковом пространстве **F** :

. (16)

#### Пример анализа информативности признаков для распознавания фонем.

Анализировались образцы фонем, выделенные при тщательном визуальном анализе сигнала «\_мама\_», . Каждый класс содержит по два образца длиной в 1024 отсчета:

.

В таблице 1 приведены спектральные составляющие сигнала, имеющие максимальную информативность. Заметим, что набор из всех спектральных составляющих имеет информативность чуть больше 0,5, которая в принципе обеспечивает распознавание последовательности состояний в потоке, но требует слишком сложного классификатора, а при вариациях распознавание будет ненадежным. Спектральные составляющие большой мощности и форманты в диапазоне 100 – 300 Гц, оказывается, не являются информативными (0,4 -0,7), как принято полагать.

Таблица 1.Спектральные составляющие с максимальной информативностью

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Частота, Гц | 70,3 | 687,5 | 406,3 | 78,1 | 3414,1 | 1718,8 | 2484,4 | 2015,6 |
| Информативность | 10,2 | 6,30 | 5,25 | 4,3 | 4,03 | 4,00 | 3,15 | 3,01 |

**Заключение**

Таким образом, если цифровая обработка сигналов разных классов позволяет построить образ с малым количеством компонент и информативностью, достаточной для классификации сигналов, или, по крайней мере, более высокой, чем информативность самого сигнала или используемых образов, то можно говорить об улучшении качества обработки сигналов. Методы оценки информативности результатов цифровой обработки сигналов позволяют выбрать алгоритмы для конкретных задач классификационного анализа сигналов до разработки целевой системы, что значительно снижает риски.

# Задание

Два класса сигналов с неодинаковыми длительностями представлены обучающей выборкой , каждая содержит по 4 сигнала.

1. Построить образы сигналов на основе статистик Wavelet-коэффициентов.
2. Оценить информативность каждого признака, включенного в образ

# Методика выполнения

1. Выполнить Wavelet-анализ каждого сигнала в программе, разработанной в предыдущей лабораторной работе
2. Вычислить статистики Wavelet-коэффициентов. Результаты представить в таблице 1.

**Таблица 1. Образы сигналов «…»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Сигнал | Имя признака 1 | Имя признака 2 | ……. | ……. | Имя признака n |
| \_u\_ | Значение признака 1 сигнала \_u\_ | ……… | …………. | ……………… | Значение признака n сигнала \_u\_ |
| ……. |  |  |  |  |  |
| u |  |  |  |  |  |
| a\_ |  |  |  |  |  |
| ………. |  |  |  |  |  |
| \_a\_ |  |  |  |  | Значение признака n сигнала \_a\_ |

1. Составить матрицу образов, дополненную характеристическим вектором (первый столбец). Компоненты образа сигнала находятся в столбцах, каждый столбец соответствует признаку. Матрицу сохранить в файле формата \*.txt. Дробная часть действительных чисел отделяется точкой, разделитель столбцов – знак табуляции.
2. Открыть файл **ICalc**.**mcd** (или ICalc.xmcd) с системой расчета информативности.
3. Указать путь к файлу с матрицей **Th** (см. Приложение 1).
4. Результат расчета информативности (см. Приложение 1) представить в таблице 2:

**Таблица 2. Информативность признаков в задаче классификации «…»**

|  |  |  |
| --- | --- | --- |
| Имя признака | Номер признака | Информативность |
| ……….. | ………. | ……… |
|  |  |  |
|  |  |  |

1. Сделать выводы

# Контрольные вопросы

1. Что представляет собой образ объекта в системах ЦОС?
2. Что представляет собой образ сигнала?
3. Коэффициент редукции сигнала.
4. Теряется ли информация при переходе от сигнала к его образу?

# Литература

1. Биргер И. А. Техническая диагностика. – М.: Машиностроение, 1978. – 240 с.
2. Колесникова С. И. Методы анализа информативности разнотипных признаков. – Вестник Томского государственного университета. – 2009. - №1(6). – С. 69 – 80.
3. Грызлова Т. П., Балыкина А. О. Система оценки информативности диагностических признаков и признаковых пространств. - Авиационно-космическая техника и технология. № 9 (86) Харьков, «ХАИ», 2011. С. 148-154

# Приложение 1. Фрагменты данных и результатов в системе ICalc.mcd

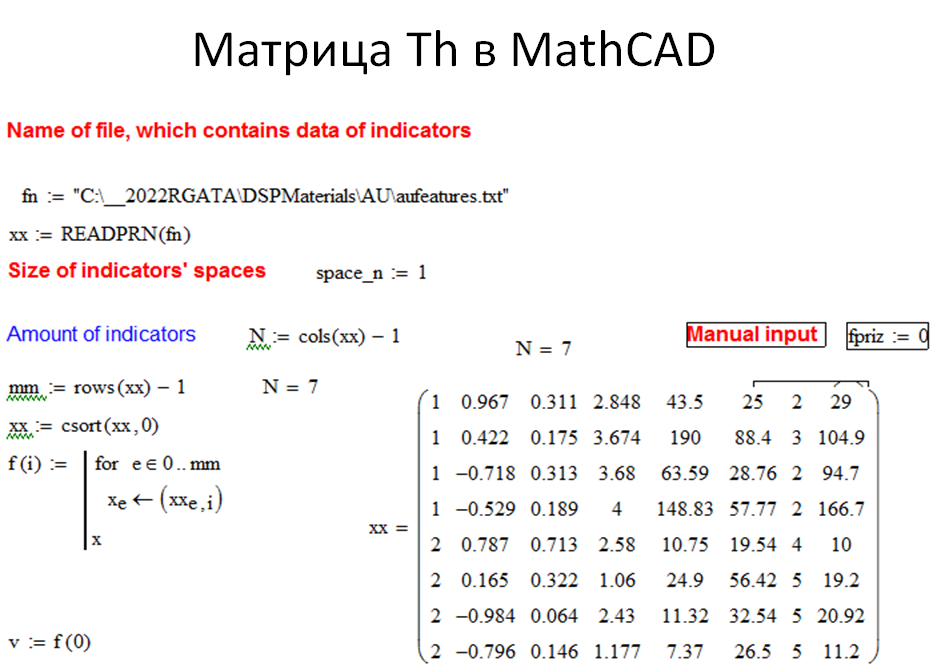


Рис. П.1. Входные данные (размерность пространства – 1)

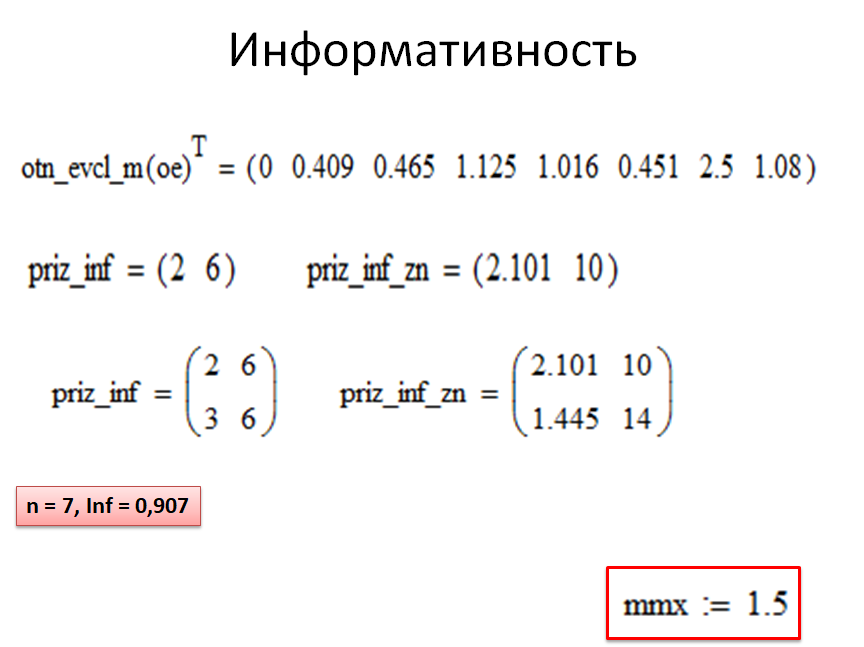


Рис. П.2. Результаты: информативность признаков, информативность пространств (2-й – 6-й признаки, 3-й – 6-й признаки). mmx – порог, если признак или пространство имеет информативность ниже порога, то они в результат выводиться не будут